University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 130

Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples

Mark J. van der Laan* Sandrine Dudoit†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper130

Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples

Mark J. van der Laan and Sandrine Dudoit

Abstract

In Part I of this article we propose a general cross-validation criterian for selecting among a collection of estimators of a particular parameter of interest based on n i.i.d. observations. It is assumed that the parameter of interest minimizes the expectation (w.r.t. to the distribution of the observed data structure) of a particular loss function of a candidate parameter value and the observed data structure, possibly indexed by a nuisance parameter. The proposed cross-validation criterian is defined as the empirical mean over the validation sample of the loss function at the parameter estimate based on the training sample, averaged over random splits of the observed sample. The cross-validation selector is now the estimator which minimizes this cross-validation criterion. We illustrate that this general methodology covers, in particular, the selection problems in the current literature, but results in a wide range of new selection methods. We prove a finite sample oracle inequality, and asymptotic optimality of the cross-validated selector under general conditions. The asymptotic optimality states that the cross-validation selector performs asymptotically exactly as well as the selector which for each given data set makes the best choice (knowing the true data generating distribution).

Our general framework allows, in particular, the situation in which the observed data structure is a censored version of the full data structure of interest, and where the parameter of interest is a parameter of the full data structure distribution. As examples of the parameter of the full data distribution we consider a density of (a

part of) the full data structure, a conditional expectation of an outcome, given explanatory variables, a marginal survival function of a failure time, and multivariate conditional expectation of an outcome vector, given covariates. In part II of this article we show that the general estimating function methodology for censored data structures as provided in van der Laan, Robins (2002) yields the wished loss functions for the selection among estimators of a full-data distribution parameter of interest based on censored data. The corresponding cross-validation selector generalizes any of the existing selection methods in regression and density estimation (including model selection) to the censored data case. Under general conditions, our optimality results now show that the corresponing cross-validation selector performs asymptotically exactly as well as the selector which for each given data set makes the best choice (knowing the true full data distribution).

In Part III of this article we propose a general estimator which is defined as follows. For a collection of subspaces and the complete parameter space, one defines an epsilon-net (i.e., a finite set of points whose epsilon-spheres cover the complete parameter space). For each epsilon and subspace one defines now a corresponding minimum cross-valided empirical risk estimator as the minimizer of cross-validated risk over the subspace-specific epsilon-net. In the special case that the loss function has no nuisance parameter, which thus covers the classical regression and density estimation cases, this epsilon and subspace specific minimum risk estimator reduces to the minimizer of the empirical risk over the corresponding epsilon-net. Finally, one selects epsilon and the subspace with the cross-validation selector. We refer to the resulting estimator as the cross-validated adaptive epsilon-net estimator. We prove an oracle inequality for this estimator which implies that the estimator minimax adaptive in the sense that it achieves the minimax optimal rate of convergence for the smallest of the guessed subspaces containing the true parameter value.

Cross-Validation for Estimator Selection

1 Stating the Selection Problem.

Let $O_1, \ldots, O_n$ be n i.i.d. observations of $O \sim P_0$, where P_0 is known to be an element of a statistical model $\mathcal{M}$. Let $\psi_0(\cdot) = \psi(\cdot \mid P_0)$ be a parameter (function) of P_0 of interest. Let the parameter set for this parameter be $\mathbf{\Psi} = \{\psi(\cdot \mid P) : P \in \mathcal{M}\}$. Let $(O, \psi) \to L(O, \psi \mid \eta_0) \in \mathbb{R}$ be a "loss function", possibly depending on a nuisance parameter $\eta_0 = \eta(P_0)$, which maps a candidate parameter value ψ and observation O into a real number, whose expectation is minimized at ψ_0:

$$\begin{aligned} \psi_0 &= \text{argmin}_{\psi \in \mathbf{\Psi}} \int L(o, \psi \mid \eta_0) dP_0(o) \\ &= \text{argmin}_{\psi \in \mathbf{\Psi}} E_0 L(O, \psi \mid \eta_0). \end{aligned} \tag{1}$$

Let P_n be the empirical distribution of $O_1, \ldots, O_n$. Let $\hat{\psi}_k(\cdot) = \psi_k(\cdot \mid P_n) \in \mathbf{\Psi}$, $k = 1, \ldots, K(n)$, be a collection of estimators (i.e., algorithms one can apply to data) of $\psi_0(\cdot)$.

The choice of loss function. Different choices of loss functions can satisfy (1). In fact, (1) can define a class of possible loss functions. Different choices of loss functions result in estimators of ψ_0 with different behavior. Consequently, the choice of loss function is an interesting issue to be addressed. We suggest the following reasonable strategy for selecting a loss function.

Firstly, among the loss functions identifying ψ_0 as the minimizer of its risk (i.e., satisfying (1), one wishes to choose a loss function which identifies the wished measure of performance/Risk

$$\tilde{\theta}(\psi \mid P_0) \equiv \int L(O, \psi \mid \eta_0) dP_0(O) \text{ for a candidate } \psi \in \mathbf{\Psi}.$$

Identifying such a function $\tilde{\theta}(\psi \mid P_0)$ on the parameter set $\mathbf{\Psi}$ does still not uniquely identify the loss function $L(O, \psi \mid \eta_0)$. Secondly, given this function $\tilde{\theta}(\psi \mid P_0)$, we now wish to choose the loss function so that for a locally consistent estimator η_n of η_0, $1/n \sum_i L(O_i, \psi \mid \eta_n)$ is a locally efficient estimator of $\tilde{\theta}(\psi \mid P_0)$. That is, let $L(O, \psi \mid \eta_0)$ be a parametrization of

the efficient influence function for the real valued parameter $\tilde{\theta}(\psi \mid P)$ in the model $\mathcal{M}$ plus the constant $\tilde{\theta}(\psi \mid P_0)$ (Bickel, Klaassen, Ritov, Wellner, 1993, van der Laan, Robins, 2002).

The choice of curve $\psi \rightarrow \tilde{\theta}(\psi \mid P_0)$ could be subject matter driven (e.g, one might prefer the squared error loss fuction above the minus logarithm loss function because of its interpretability), but this choice can also be driven by mathematical properties of this curve in a neighborhood of ψ_0 such as its derivatives at ψ_0 in various allowed directions.

The Selection Problem: One of the most important statistical problems is the selection of an estimator among a class of candidate estimators of a common parameter of interest. Such a selection problem can only be properly defined by defining a distance or dissimilarity between a candidate estimator and the parameter of interest. We choose as dissimilarity

$$d_n(\hat{\psi}_k, \psi_0) \equiv \int \{L(o, \psi_k(\cdot \mid P_n) \mid \eta_0) - L(o, \psi_0 \mid \eta_0)\} \, dP_0(o).$$

Let $d(\psi, \psi_0) \equiv \int L(o, \psi \mid \eta_0) - L(o, \psi_0 \mid \eta_0) dP_0(o)$ so that $d_n(\hat{\psi}_k, \psi_0) = d(\psi(\cdot \mid P_n), \psi_0)$. We note that for all $\psi \in \mathbf{\Psi}$ $d(\psi, \psi_0) \geq 0$, and if the mimimum ψ_0 is unique, then $d(\psi, \psi_0) = 0$ if and only if $\psi = \psi_0$. As in the prediction literature, we will refer to

$$\tilde{\theta}_n(k) \equiv \int L(o, \psi_k(\cdot \mid P_n) \mid \eta_0) dP_0(o)$$

as the "conditional risk" of the estimator $\psi_k(\cdot \mid P_n)$. Throughout this paper we will use the analogous terminology as in the prediction literature for the quantaties of interest.

Let

$$\begin{aligned} \tilde{k}_n &\equiv \text{argmin}_k d_n(\hat{\psi}_k, \psi_0) \\ &= \text{argmin}_k \int L(o, \psi_k(\cdot \mid P_n) \mid \eta_0) dP_0(o) \\ &= \text{argmin}_k \tilde{\theta}_n(k) \end{aligned} \tag{2}$$

be the **optimal benchmark selector** which chooses for each given data set $O_1, \ldots, O_n$ the estimator with minimal dissimilarity to the true parameter value ψ_0, or equivalently, with minimal conditional risk $\tilde{\theta}_n(k)$. If the mimimum is not unique, then the argmin is defined as the smallest k achieving the minimum. Note that this benchmark selector depends on the unknown data generating distribution P_0.

It follows that a data adaptive selector $\hat{k} = \hat{k}(P_n)$ is **asymptotically equivalent** with this oracle benchmark selector $\tilde{k}_n$ if

$$\frac{d_n(\hat{\psi}_{\hat{k}}, \psi_0)}{d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)} \to 1 \text{ in probability.}$$

Our general theorems later in this chapter will precisely establish this result for the cross-validation selector defined in the next section, under general conditions.

2 The cross-validation selector.

Define random vector $S_n \in \{0,1\}^n$ for splitting the sample into a *validation* and a *training* sample.

$$S_{n,i} = \begin{cases} 0 & \text{if i-th observation is in the training sample} \\ 1 & \text{if i-th observation is in the validation sample} \end{cases}$$

Different choices of distributions for S_n cover many types of cross-validation including $V-$ fold cross-validation and Monte Carlo cross validation. For example, 5-fold cross-validation corresponds with 5 possible outcomes of S_n each having equal probability. Below, we will discuss such choices in more detail.

Let $p = n_1/n$ be the proportion constituting the validation sample. Let P^0_{n,S_n} and P^1_{n,S_n} be the empirical distributions of the training and validation sample, respectively. The cross-validation selector is defined by:

$$\begin{aligned} \hat{k} &\equiv \text{argmin}_k E_{S_n} \int L(o, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) dP^1_{n,S_n}(o) \qquad (3) \\ &= \text{argmin}_k E_{S_n} \frac{1}{np} \sum_{i=1}^{n} I(S_n(i) = 1) L(O_i, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}). \end{aligned}$$

Here $E_{S_n} \int L(o, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) dP^1_{n,S_n}(o)$ represents an estimator of the true conditinal risk $\tilde{\theta}_n(k)$. Given any estimator $\psi(\cdot \mid P_n)$, one can view this risk estimate $E_{S_n} \int L(o, \psi(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) dP^1_{n,S_n}(o)$ as an performance assessment of the estimator $\psi(\cdot \mid P_n)$.

The proportion of observations in the validation sets, $p = \sum_i S_{n,i}/n$, is typically a pre-specified parameter of the CV procedure, with $p \in (0,1)$. When needed, we will use the notation p_n to emphasize the dependence of p on the sample size n.

2.1 Possible cross-validation schemes.

We will now discuss the possible choices for the distribution of S_n.

Monte Carlo cross-validation: In *Monte Carlo cross-validation*, the learning set is repeatedly and randomly divided into two sets, a training set of $n_0 = n(1-p)$ observations and a validation set of $n_1 = np$ observations. A common choice for p in the machine learning literature is 10% (Breiman, 1998). The corresponding distribution of S_n places mass $1/\binom{n}{np}$ on each binary vector $s_n = (s_{n,1}, \ldots, s_{n,n})$ such that $\sum_i s_{n,i} = np$. In practice, the support of the distribution of S_n can be very large and one approximates the expected value over S_n by an empirical average based on a random sample of S_n's.

V-fold cross-validation: In *V-fold cross-validation*, the learning set $\mathcal{L} = \{O_1, \ldots, O_n\}$ is randomly divided into V mutually exclusive and exhaustive sets, $\mathcal{L}_v$, $v = 1, \ldots, V$, of as nearly equal size as possible. Estimators are built on training sets $\mathcal{L} - \mathcal{L}_v$, risk estimates are computed for the validation sets $\mathcal{L}_v$, and averaged over v. V-fold CV amounts to using a random vector S_n with a distribution that places mass $1/V$ on each of the V binary vectors s_n^v, $v = 1, \ldots, V$, defined as follows. Let $n_V = \lfloor n/V \rfloor$ denote the integer part, or floor, of n/V. Then, for $v = 1, \ldots, V-1$, let $s_{n,i}^v = 1$ for $i = 1 + (v-1)n_V, \ldots, vn_V$ and 0 otherwise. For $v = V$, let $s_{n,i}^V = 1$ for $i = 1 + (V-1)n_V, \ldots, n$ and 0 otherwise. The proportion p of observations in the validation sets is approximately $1/V$.

Leave-one-out cross-validation: A commonly used form of cross-validation is *leave-one-out cross-validation* (LOOCV), where $V = n$ and $p_n = 1/n$. In LOOCV, each observation in the learning set is used in turn as the validation set and the remaining $n-1$ observations are used as the training set. The corresponding distribution of S_n places mass $1/n$ on each binary vector $s_n = (s_{n,1}, \ldots, s_{n,n})$ such that $\sum_i s_{n,i} = 1$. Our finite sample and asymptotic results in this chapter require that $np_n \to \infty$; this is not the case for LOOCV.

Intuitively, there is a *bias-variance trade-off* in the selection of p. Large p's typically produce estimators of the conditional risk $\tilde{\theta}_n$ with a large bias, but a small variance. In particular, LOOCV, with $p = 1/n$, often results in low bias but high variance estimators. The simulation studies

in Breiman and Spector (1992) and Breiman (1996a) for model selection in prediction show that leave-one-out cross-validation is inferior to leave-many-out cross-validation (e.g., $V = 10$-fold CV). In particular, LOOCV is found to behave poorly in selection from an unstable sequence of predictors.

Bootstrap cross-validation: A number of cross-validation procedures based on bootstrap samples have been proposed for estimating prediction error (Ambroise and McLachlan, 2002; Efron and Tibshirani, 1993). The standard *leave-one-out bootstrap* procedure, $B1$, can be viewed as producing training sets that are random samples of size n drawn *with replacement* from the learning set. For each bootstrap sample, about one-third ($(1 - 1/n)^n \approx e^{-1} \approx .368$) of the cases are left out; these observations form the validation set. The definition of the random vector S_n needs to be modified for bootstrap-based CV to account for multiple occurrences of the same observation in the training sets. This can be done by allowing weights in the empirical distribution P^0_{n,S_n}. In this setting, one could define $s_{n,i}$ as the number of occurrences of observation (X_i, Y_i) in the training set, so that $S_n \in \{0, \ldots, n\}^n$ and there are n^n possible random vectors S_n. In practice, one approximates the expected value over S_n by an empirical average based on a random sample of S_n's. For bootstrap-based CV, the proportion of observations in the validation sets, $p_n = \sum_i I(S_{n,i} = 0)/n$, is a random variable, with $E[p_n] = (1 - 1/n)^n \approx e^{-1} \approx .368$ (note that $S_{n,i} = 0$ now correspond to validation set observations). Shortcomings of this cross-validation scheme include the occurrence of ties in the training sets and the lack of control over p_n. The *.632 bootstrap estimator* is a convex combination of this bootstrap cross-validation risk estimator with weight 0.632 and the substitution estimator with weight 1-0.632. Here the factor .632 corresponds to the expected proportion of learning set observations included in the bootstrap training samples (Chapter 17 in Efron and Tibshirani (1993)).

3 Examples.

In this section we present six examples to which we can apply the general cross-validation selection procedure. These examples cover, in particular,

the current literature, but also completely new important selection methods. This set of examples is not meant to be exhaustive. In Chapter 2 we will show that our cross-validation methodology covers, in particular, the generalization of a cross-validation selection method based on a particular full data structure X (e.g., predictor selection, model selection, etc) to any censored data structure $O = \Phi(C, X)$ for a known many to one-mapping Φ and censoring variable C.

Example 1 (Predictor Selection) We observe n i.i.d. observations of $O = (Y, W) \sim P_0$, Y is an outcome, W is a vector of covariates. Let $\psi_0(W) = E_0(Y \mid W)$ be the parameter of interest. If we define

$$L(Y, W, \psi) = (Y - \psi(W))^2$$

as the quadratic loss function, then $\psi_0 = \text{argmin}_\psi E_0 L(Y, W, \psi)$. Given candidate predictors $\hat{\psi}_k = \psi_k(\cdot \mid P_n)$, we have

$$d_n(\hat{\psi}_k, \psi_0) = \int (\psi_k(w \mid P_n) - \psi_0(w))^2 \, dF_W(w).$$

The cross-validated selector is given by:

$$\hat{k} \quad = \quad \text{argmin}_k E_{S_n} \sum_{i:S_n(i)=1)} (Y_i - \psi_k(W_i \mid P^0_{n,S_n}))^2$$

As discussed by Breiman (1992) in the context of dimensionality selection in regression, criteria such as Mallow's C_p, Akaike information's criterion (AIC), and the Bayesian information criterion (BIC), do not account for the data-driven selection of the sequence of models and thus provide biased assessment of prediction error in finite sample situations. Instead, risk estimation methods based on sample reuse have been favored. The main procedures include: leave-one-out cross-validation, V-fold cross-validation, Monte Carlo cross-validation, and the bootstrap (Chapter 3 in Breiman et al. (1984a), Breiman and Spector (1992), Breiman (1996a), Breiman (1996b), Chapter 17 in Efron and Tibshirani (1993), Chapters 7 and 8 in Györfi et al. (2002), Chapter 7 in Hastie et al. (2001), Chapter 3 in Ripley (1996), Stone (1974b), Stone (1977)).

Thus, a variety of cross-validation procedures are available for estimating the risk of a predictor. A natural question then concerns the distributional properties of the resulting risk estimators, i.e., their performance as

estimators of generalization error, their performance in terms of identifying a good predictor (model selection), and also the impact of the particular cross-validation procedure (e.g., the choice of V in V-fold cross-validation, the use of V-fold vs. Monte Carlo cross-validation). Aside from empirical assessment of different estimation procedures, previous theoretical work has focused primarily on the distributional properties of leave-one-out cross-validation (Stone, 1974b, 1977).

There is a rich literature on leave-one-out cross-validation in nonparametric univariate regression. For example, Silerman (1984) proposes a fast approximation of the leave-one out cross-validation method in spline regression. We refer to Härdle (1993) for an overview on the leave-one-out cross-validation method in kernel regression. In particular, Härdle and Marron (1985a) and Härdle and Marron (1985b) establish an asymptotic optimality result for leave-one-out cross-validation for choosing the smoothing parameter in nonparametric kernel regression (see page 158, Härdle (1993)).

Györfi et al. (2002) recently established a finite sample result for the single-split cross-validation selector for the squared error loss function. Their theorem was generalized in Dudoit, van der Laan (2003) to general cross-validation schemes and a general class of loss functions (and some corrections were made). Dudoit, van der Laan (2003) examine the distributional properties of cross-validated risk estimators in the context of both predictor selection and predictor performance assessment for a general class of loss functions.

Application of our general theorems 1 and 2 to this example results in the same results as established in Dudoit, van der Laan (2003) $\square$

Example 2 (Density estimator selection) We observe n i.i.d. observations on $O \sim f_0 \equiv \frac{dP_0}{d\mu}$, where μ is a dominating measure of the data generating distribution P_0. Let the parameter of interest $\psi_0 = f_0$ be the density itself. If we define

$$L(O, g) = -\log(g(O)),$$

then $f_0 = \text{argmin}_g E_0 L(0, g)$. Given candidate density estimators $\hat{\psi}_k = f_k(\cdot \mid P_n)$ of $\psi_0 = f_0$, we have

$$d_n(\hat{\psi}_k, \psi_0) = \int \log\left(\frac{f_k(o \mid P_n)}{f_0(o)}\right) f_0(o) d\mu(o),$$

that is, $d_n(\hat{\psi}_k, \psi_0)$ is the Kullback-Leibler distance between $f_k(\cdot \mid P_n)$ and f_0. For example, $f_k(\cdot \mid P_n)$ can be the maximum likelihood estimator of f according to a model $\mathcal{M}_k$, that is,

$$f_k(\cdot \mid P_n) = \text{argmax}^{-1}_{f \in \mathcal{M}_k} \int \log(f(x)) dP_n(x).$$

The cross-validation selector is given by

$$\hat{k} \;=\; \text{argmin}_k E_{S_n} \sum_{i:S_n(i)=1} -\log\left(f_k(O_i \mid P^0_{n,S_n})\right)$$

Density estimation arises in important and common problems in the statistical literature. Bandwidth selection in kernel density estimation, selecting the number of components in mixture models, and variable selection in regression (e.g., logistic and linear regression with normal error), are three examples that involve density estimation.

Leave-one-out likelihood cross-validation in density estimation is discussed in Silverman (1986) who refers to Stone (1974a) and Geisser (1975) for its general applicability to model fitting as well. Silverman (1986) refers to Scott and Factor (1981) to indicate that for densities with infinite support this leave-one-out likelihood cross-validation method for bandwidth selection in density estimation is sensitive to outliers, and to Schuster and Gregory (1981) to point out that leave-one-out cross-validation can result, in fact, into inconsistent density estimators under non-pathological conditions. Stone (1984) provides an asymptotically optimal bandwidth selection rule for kernel density estimation, which has a leave-one out cross-validation interpretation.

Recent work on (V-fold or Monte-Carlo) cross-validated likelihood methods for choosing the number of components in mixture models is found in Smyth (2000) and Pavlic and van der Laan (2003). In particular, the simulation studies of Pavlic and van der Laan (2003) showed that likelihood based cross-validation performed well compared to common approaches based on validity functionals such as Akaike's information criterion (Akaike (1973), Bozdogan (2000)), Bayesian Information criterion BIC (Schwartz (1978)) or Minimum description length (Rissanen (1978), see Hansen and Yu (2001), for an overview) and ICOMP (Bozdogan (1993)).

Likelihood based cross-validation covers in particular squared error-loss cross-validation for prediction. Specifically, let $\mathcal{M}_k$ be a regression model $Y = \mu_k(Z) + N(0, \sigma^2)$, with μ_k ranging over a family of curves indexed by

k, and let $f_k(X \mid P_n)$ be the corresponding least squares estimator (i.e., maximum likelihood estimator) .

Van der Laan, Dudoit, Keles (2003) study this general likelihood based cross-validation selector and establish, under general conditions on P_0, that the cross-validation selector $\hat{k}$ for k is asymptotically optimal, in the sense that it performs as well as the optimal benchmark selector $\tilde{k}_n$ (2) based on the true data generating distribution P_0. They also illustrate this asymptotic result and the practical performance of likelihood based cross-validation for the purpose of bandwidth selection in density estimation with a simulation study. Application of our general Theorem 1 to this example yields the same results as in van der Laan, Dudoit, Keles (2003). $\square$

Example 3 (Multivariate predictor selection) Let $O = (Y = (Y_1, \ldots, Y_l), W) \sim P_0$, where Y is a multivariate random outcome vector and W a vector of covariates. Let $\psi_0(W) \equiv E(Y \mid W) = (E(Y_1 \mid W), \ldots, E(Y_l \mid W))$ be the multivariate conditional expectation of Y, given W. For a candidate multivariate predictor $\psi(W)$, we define

$$L(O, \psi \mid \eta_0) \equiv (Y - \psi(W))^\top \eta_0(W)(Y - \psi(W)),$$

where η_0 is a symmetric $l \times l$-matrix function of W. If η_0 is a user supplied known matrix, then it is not a nuisance parameter and we can denote the loss function with $L(O, \psi)$. However, η_0 can also denote the limit of an estimator of an unknown matrix such as

$$\left[E_0\left(\{Y - E_0(Y \mid W)\}\{Y - E_0(Y \mid W)\}^\top \mid W\right)\right]^{-1}.$$

In this case η_0 denotes a nuisance parameter which needs to be estimated from the data. For any symmetric matrix function $\eta(W)$ we have

$$\psi_0 = \text{argmin}_\psi E_0 L(O, \psi \mid \eta). \tag{4}$$

Given candidate estimators $\psi_k(\cdot \mid P_n)$ of ψ_0, $k = 1, \ldots, K(n)$, and an estimator $\eta(P_n)$ (e.g., an estimate of the inverse of the conditional covariance matrix according to a working model such as the independence working model) of η_0 , the cross-validation selector $\hat{k}$ is given by:

$$\begin{aligned}
\hat{k} &= \text{argmin}_k E_{S_n} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta(P^0_{n,S_n})) dP^1_{n,S_n}(O) \\
&= \text{argmin}_k E_{S_n} \frac{1}{np} \sum_{i=1}^{n} \{I(S_n(i) = 1) \\
&\quad (Y_i - \psi_k(W_i \mid P^0_{n,S_n}))^\top \eta(W_i \mid P^0_{n,S_n})(Y_i - \psi_k(W_i \mid P^0_{n,S_n}))\Big\}.
\end{aligned}$$

We note that

$$(Y-\psi_k(W \mid P^0_{n,S_n}))^\top \eta_0(W)(Y-\psi_k(W \mid P^0_{n,S_n})) = \| \eta_0^{0.5}(W)(Y-\psi_k(W \mid P^0_{n,S_n})) \|^2,$$

where $\| x \| = \sqrt{\sum_{j=1}^l x_j^2}$ is the euclidean norm in $\mathbb{R}^l$, and $\eta_0^{0.5}$ is the square root of η_0. This shows that

$$d_n(\hat{\psi}_k), \psi_0) = \int \| \eta_0^{0.5} (\psi_k(W \mid P_n) - \psi_0(W)) \|^2 \, dF_0(W).$$

Application of our general Theorem 1 to this example results in a finite sample result and asymptotic optimality. □

Example 4 (Counterfactual predictor selection in causal inference) Let $X = ((Y_a, a \in \mathcal{A}), W) \sim F_{X,0}$ be the full data structure of interest, where W denotes baseline covariates and Y_a denotes the outcome on a subject if the subject would have taken treatment a. Such potential outcomes Y_a are called counterfactuals (e.g Rubin, 1978). Let A be a random variable with conditional probability distribution $g_0(a \mid X) \equiv P(A = a \mid X)$, which denotes the treatment the subject actually took. We will only observe the outcome indexed by the treatment the subject took. Thus we observe n i.i.d. observations of $O = (A, Y_A, W)$. We assume that treatment is randomized within strata of W: $g_0(a \mid X) = g_0(a \mid W)$ for all $a \in \mathcal{A}$. We have that the distribution $P_0 = P_{F_{X0},g_0}$ is indexed by the full data distribution F_{X0} and the conditional density g_0. Suppose that the parameter of interest is $\psi_0(a, V) = E(Y_a \mid V)$, that is, we want to estimate the multivariate regression of the vector $(Y_a : a \in \mathcal{A})$ of potential outcomes on V. If we would observe X, then this would be the same problem as covered in the previous multivariate prediction example with η_0 being the identity matrix. Thus, in the case that we would observe the full data structure we would use as loss function

$$L(X, \psi) = \sum_{a \in \mathcal{A}} (Y_a - \psi(a, V))^2.$$

Indeed we have

$$\psi_0 = \text{argmin}_\psi E_{F_{X0}} L(X, \psi).$$

In this example, we only observe one of the outcomes for each subject. We will choose as loss function the IPTW or double robust mapping applied to

this full data loss function $L(X,\psi)$ (van der Laan and Robins (2002), Section 6.3):

$$\begin{aligned} L(O,\psi \mid \eta_0) &= IC(O \mid Q_0, g_0, L(\cdot,\psi)) \\ &\equiv \frac{(Y-\psi(A,V))^2}{g_0(A \mid W)} - \frac{1}{g_0(A \mid W)} E_0((Y-\psi(A,V))^2 \mid A,W) \\ &\quad + \sum_{a \in \mathcal{A}} E_0((Y-\psi(A,V))^2 \mid A=a,W). \end{aligned}$$

Here $\eta_0 = (g_0, Q_0)$ and $Q_0(A,W) = (E(Y \mid A,W), E(Y^2 \mid A,W))$. Note that the conditional expectations are indeed identified by these first two conditional moments of the conditional distribution of Y, given W. It can be verified (van der Laan and Robins (2002), Section 6.3) that for any g_1 satisfing the so called experimental treatment assignment assumption (ETA), that is, $\min_{a \in \mathcal{A}} g_1(a \mid W) > 0$ P_0-a.e., we have

$$E_{P_0} IC(O \mid Q_1, g_1, L(\cdot,\psi)) = E_{F_{X_0}} L(X,\psi) \text{ if either } g_1 = g_0 \text{ or } Q_1 = Q_0.$$

This identity is referred to as double robustness of the estimating function $IC(O \mid Q_0, g_0, L(\cdot \mid \psi))$ for $E_0 L(X,\psi)$ w.r.t. misspecification of Q_0, g_0. So we can choose η_0 to be any element in $\Gamma(P_0) \equiv \{(Q,g) : Q = Q_0 \text{ or } g = g_0\}$, where g ranges over conditional distributions satisfying ETA: that is,

$$\psi_0 = \text{argmin}_\psi E_{P_0} L(O,\psi \mid \eta_0) \text{ if } \eta_0 \in \Gamma(P_0).$$

For any $\eta_0 \in \Gamma(P_0)$, we have

$$d_n(\hat{\psi}_k, \psi_0) = \sum_{a \in \mathcal{A}} \int (\psi_k(a,V \mid P_n) - \psi_0(a,V))^2 \, dF_0(V).$$

Let $\psi_k(\cdot \mid P_n)$ be an estimator of ψ_0, $k = 1,\ldots,n$, based on n i.i.d. observations $O_1,\ldots,O_n$. For example, $\psi_k(\cdot \mid P_n)$ is an Inverse probability of treatment weighted (IPTW) estimator or double robust IPTW estimator according to a k-specific marginal structural model $E(Y_a \mid V) = m_k(a,V \mid \beta_k)$ (see Robins, 2002, and van der Laan, Robins, 2002). Given estimators Q_n, g_n of Q_0, g_0, our cross-validation selector $\hat{k}$ is given by:

$$\hat{k} = \text{argmin}_k E_{S_n} \sum_{i=1}^{n} I(S_n(i)=1) IC(O_i \mid Q^0_{n,S_n}, g^0_{n,S_n}, L(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}))),$$

where Q^0_{n,S_n}, g^0_{n,S_n} are the estimators applied to the training empirical distribution P^0_{n,S_n}.

This is a new selection method in causal inference and application of our general Theorem 1 yields a finite sample result and asymptotic optimality for this selector $\hat{k}$ under general conditions. One of the main conditions is that either g^0_{n,S_n} is consistent for g_0 or Q^0_{n,S_n} is consistent for Q_0. $\square$

Example 5 (Survival predictor selection based on right-censored data) Let $X(t)$, $t \geq 0$, be a time-dependent process, which includes as component $R(t) = I(T \leq t)$, where T is a survival time. Let $X = \bar{X}(T) \equiv \{X(t) : t \leq T\}$ be the full-data structure of interest. Let $W = X(0)$ denote the baseline covariates measured at baseline. The distribution of X will be denoted with $F_{X,0}$. Let C be a right-censoring time so that the observed data structure is given byL

$$O = (\tilde{T} = \min(T, C), \Delta = I(\tilde{T} = T), \bar{X}(\tilde{T}).$$

We will assume that the conditional distribution $G_0(\cdot \mid X)$ of C, given X, satisfies coarsening at random: that is, for $t < T$,

$$\lambda_C(t \mid X) = m(t, \bar{X}(t)) \text{ for some measurable function } m.$$

Here $\lambda_C(t \mid X)$ denotes the discrete or continuous conditional hazard of C, given X. If $X = (T, W)$ does not include time-dependent covariates, then CAR is equivalent with assuming that C is conditionally independent of T, given W. Under CAR, the density of P_0 factors into a F_{X0}-part and G_0-part (Gill, van der Laan, Robins, 1997). The F_{X0}-part of the density will be denoted with Q_{X0}.

We observe n i.i.d. observations $O_1, \ldots, O_n$ of the the right-censored data structure defined by $O = (\min(Y, C), \Delta = I(Y \leq C), W) \sim P_0 = P_{F_{X,0},G_0}$. Let $\psi_0(W) = E_0(Y \mid W)$ be the parameter of interest, where $Y = \log(T)$. The corresponding full-data loss function is given by $L(X, \psi) = (Y - \psi(W))^2$:

$$\psi_0 = \text{argmin}_\psi \int L(x, \psi) dF_{X,0}(x).$$

Suppose that

$$\bar{G}_0(T \mid X) \equiv P(C > t \mid X)|_{t=T} > \delta > 0, \ F_{X0}\text{-a.e. for some } \delta > 0.$$

Then it follows that

$$\begin{aligned}\psi_0(W) &= \text{argmin}_\psi \int L(x,\psi) dF_{X,0}(x) \\ &= \text{argmin}_\psi E_{F_{X,0}}(Y - \psi(W))^2 \\ &= \text{argmin}_\psi E_{P_0}\left\{ L(X,\psi) \frac{\Delta}{\bar{G}(T \mid X)} \right\},\end{aligned}$$

Define

$$IC(O \mid G, D) = D(X) \frac{\Delta}{\bar{G}(T \mid X)}$$

as the so called inverse probability of censoring weighted mapping from full data functions $D(X)$ to observed data functions as provided in Robins and Rotnitzky (1992). If we choose

$$L(O, \psi \mid \eta_0 = G_0) = IC(O \mid G_0, L(\cdot, \psi)),$$

then $\psi_0 = \text{argmin}_\psi E_0 L(O, \psi \mid \eta_0)$. A more sophisticated choice is defined by the so called double robust mapping from full data functions $D(X)$ to observed data functions (van der Laan, Robins, 2002, chapter 3), as defined by:

$$\begin{aligned}L(O, \psi \mid \eta = Q_{X0}, G_0) &= IC(O \mid Q_{X0}, G_0, L(\cdot, \psi)) \\ &\equiv IC_0(O \mid G_0, L(\cdot, \psi)) \qquad (5) \\ &\quad + \int E_{Q_{X0}, G_0} \left(IC_0(O \mid G_0, L(\cdot, \psi)) \mid \bar{X}(u), \tilde{T} \geq u \right) dM_{G_0}(u) \qquad (6)\end{aligned}$$

where

$$dM_{G_0}(u) = I(\tilde{T} \in du, \Delta = 0) - I(\tilde{T} \geq u) \frac{dG_0(u \mid X)}{\bar{G}_0(u- \mid X)}.$$

Note that $E_{Q_{X0}, G_0} = E_{P_0}$.

Given candidate predictors $\hat{\psi}_k = \psi_k(\cdot \mid P_n)$ (e.g. predictors based on Cox-proportional hazard model fits, or linear regression fits), the corresponding distance is given by:

$$\begin{aligned}d_n(\hat{\psi}_k, \psi_0) &= \int L(O, \hat{\psi}_k \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\ &= \int \left(\psi_k(w \mid P_n) - \psi_0(w)\right)^2 dF_W(w).\end{aligned}$$

For the IPCW-choice of loss function we have:

$$\hat{k} \;=\; \text{argmin}_k E_{S_n} \sum_{i:S_n(i)=1} (Y_i - \psi_k(W_i \mid P^0_{n,S_n}))^2 \frac{\Delta_i}{\bar{G}^0_{n,S_n}(T_i \mid W_i)},$$

where $\bar{G}^0_{n,S_n}$ is an estimator of the survivor function $\bar{G}_0(\cdot \mid X)$ based on the training sample. We note that this cross-validation selector reduces to the standard cross-validation selector in Example 1 in case there is no censoring. The asymptotic validity of this selector relies on the consistency of the estimator of the survivor function $\bar{G}_0$. If one uses the double robust loss function, then the asymptotic validity of the corresponding selector only relies on the consistency of either $\bar{G}_n$ or of the estimator Q_{Xn} of Q_{X0}.

Nonparametric and semiparametric regression methods are among the most popular methods for analyzing censored survival data. Recently, nonparametric alternatives to Cox proportional hazards model have gained importance. These methods adapt well-known techniques of regression analysis to the analysis of censored survival data. In general, the available methods propose a sequence of regression models of increasing complexity, typically determined by a stepwise feature selection method. These sequence of models are referred to as sieves. The final choice of the model is determined with a particular model selection criteria among these sieves. Although the actual regression methodology for analyzing censored data has been extensively studied, the problem of selecting the best model or predictor among a given set of sieves has not gained much attention. We firstly give a brief overview of the available nonparametric and semiparametric regression methods for censored survival data and the model selection methods used by them.
Some of the most commonly used regression methods for censored survival data are based on splines or partitioning trees. In particular, Hastie and Tibshirani (1990b) and Hastie and Tibshirani (1990a) use additive Cox-proportional hazards models that model covariate effects with smoothing splines. Kooperberg et al. (1995) follow a polynomial spline approach and propose a sieve of multiplicative intensity models for the hazard of survival which allows interaction effects between covariates and with time. Model selection techniques such as AIC Akaike (1973); Bozdogan (2000), BIC Schwartz (1978) are used to data adaptively select the best model. Several extensions of (classification and) regression trees, CART, Breiman et al. (1984b) have also been proposed for censored survival data. These are sometimes referred to as survival trees and can roughly be divided into two categories. Methods in the first category use a *within node homogeneity measure*. Examples

of such approaches are provided in Gordon and Olshen (1985), Davis and Anderson (1989), and Leblanc and Crowley (1992). To be more specific, for instance, Davis and Anderson (1989) use the negative log-likelihood of an exponential model for each node as a measure of split node homogeneity and the squared difference of the parent node log-likelihood and a weighted sum of child node log-likelihoods as the split function. Methods in the second category, first proposed by Segal (1988), use a *between node homogeneity/heterogeneity measure* and a split function based on the two sample log rank test statistic.

In essence, these methods bypass evaluation of risk of a given predictor based on censored data by replacing the least squares split functions utilized by CART in the uncensored continuous outcome setting with alternatives. In summary, the available spline-based regression methods for censored survival data use AIC, BIC or variants thereof for model selection, whereas the tree-based methods replace the least squares split criterion with an alternative split criterion that can easily handle censored data.
In prediction and model selection problems with uncensored data, resampling-based risk estimators that are obtained by $V-$fold cross-validation or Monte carlo cross validation (repeated sample splitting) are commonly used Breiman et al. (1984b); Burman (1989); Shao (1993); Zhang (1993): see Example 1. The performance assessment of a given predictor with uncensored outcome is achieved by estimating its risk with respect to a user supplied loss function with the empirical mean of the corresponding loss function over a (independent) validation sample. Contrary to the prediction and model selection literature with the uncensored outcome, we observe that, in general, the literature on censored survival data does not seem to propose means for assessing the performance of predictors. Apparently, the problem was how to calculate risk when the outcome is subject to censoring.

Keleş et al. (2002) introduce the selector $\hat{k}$ (and the one based on double robust inverse probability of censoring weighted risk estimates) as a model selection method to select among such predictors of censored survival outcomes. We note that the cross-validated risk criteria minimized by the selector $\hat{k}$ generalizes the cross-validated risk criteria used with uncensored data to censored data. As shown in Keleş et al. (2002), under general conditions, the selector $\hat{k}$ defined above is asymptotically equivalent with the optimal benchmark selector for predictor selection in regression problems with censored outcome. In this method the risk of a given predictor based on the training sample is considered a full data parameter in a censored data model.

Subsequently, we utilize inverse probability of censoring weighted and doubly robust locally efficient estimation methods for estimating this parameter based on the validation sample as presented in Robins and Rotnitzky (1992); Robins et al. (2000); Robins and Rotnitzky (2001); van der Laan and Robins (2002). This risk estimation method also handles informative censoring by incorporating covariates in the model for the censoring mechanism. If one uses the inverse probability censoring weighted risk estimator, then the performance of a given predictor is assessed consistently as long as the censoring mechanism is estimated consistently. Application of our general Theorem 1 to this example yields a similar asymptotic result as the one established in Keleş et al. (2002), but it also provides a finite sample inequality of interest. □

Example 6 (Survival function estimator selection. Consider the same right-censored data structure $O = (\tilde{T} = \min(T, C), \Delta = I(T \leq C), \bar{X}(\tilde{T}) \sim P_0 = P_{F_{X,0},G_0}$, as defined in the previous example. As in the previous example, we leave the full data model unspecified and assume CAR on G_0. Suppose now that the parameter of interest is the survival function $\psi_0 = S_0(t) \equiv P(T \geq t)$ of T at a particular time point t.

In this case the corresponding full-data loss function is given by

$$L(T, \psi) = (I(T > t) - \psi)^2.$$

Then it follows that, if $\bar{G}_0(t \mid X) > 0$, F_X-a.e., then

$$\begin{aligned} \psi_0 &= \text{argmin}_\psi E_0 L(T, \psi) \\ &= \text{argmin}_\psi E_{P_0} \left\{ L(T, \psi) \frac{I(C > \min(t, T))}{\bar{G}(\min(T, t) \mid X)} \right\} \\ &\equiv \text{argmin}_\psi E_{P_0} IC_0[O \mid G_0, L(\cdot, \psi)]. \end{aligned}$$

Thus, if we choose $L(O, \psi \mid \eta_0 = G_0) = IC_0(O \mid G_0, (\cdot, \psi))$, then $\psi_0 = \text{argmin}_\psi E_0 L(O, \psi \mid \eta_0)$. Alternatively, we can choose the corresponding double robust loss function:

$$L(O, \psi \mid \eta_0 = (Q_{X0}, G_0)) = IC(O \mid Q_{X0}, G_0, L(\cdot, \psi)),$$

as defined in (6).

Given candidate predictors $\hat{\psi}_k = \psi_k(P_n)$, the corresponding distance is given by:

$$\begin{aligned} d_n(\hat{\psi}_k, \psi_0) &= \int L(O, \hat{\psi}_k \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\ &= (\psi_k(P_n) - \psi_0)^2 . \end{aligned}$$

For this IPCW-choice of loss function we have:

$$\hat{k} = \text{argmin}_k E_{S_n} \sum_{i:S_n(i)=1} (I(T_i > t) - \psi_k(P^0_{n,S_n}))^2 \frac{\Delta_i}{\bar{G}^0_{n,S_n}(T_i \mid W_i)}.$$

One crucial property of this cross-validation selector $\hat{k}$ is that it aims to minimize $d_n(\hat{\psi}_k, \psi_0)$ and thus it aims to choose the estimator which is closest to the true parameter value. For example, suppose that $\hat{\psi}_k$ are estimators based on different Cox-proportional hazards models or linear regression models. If one would apply standard model selection methodology such as AIC one would aim to choose the model which best fits the density of the data, while our cross-validator selector chooses the model which gives the best estimate of the parameter of interest. This selection method is new and has not been handled in the current literature. Application of our general Theorem 1 yields a finite sample result and asymptotic equivalence with the optimal benchmark selector $\tilde{k}_n$. □

Example 7 (Density or hazard estimator selection with right-censored data) Consider the same right-censored data structure $O = (\tilde{T} = \min(T, C), \Delta = I(T \leq C), \bar{X}(\tilde{T}) \sim P_0 = P_{F_{X,0},G_0}$, as defined in the previous Example 5. As in this example, we leave the full data model unspecified and assume CAR on G_0. Suppose now that the parameter of interest is the full data density $\psi_0(X_l, X_r) = f_{X0}(X_l \mid X_r)$, where X_l, X_r are components of the full-data structure X. For example, $X_l = T$ and $X_r = W$ so that ψ_0 denotes the conditional density of survival T, given the baseline covariates W.

In this case the corresponding full-data loss function is given by

$$L(X, \psi) = -\log \psi(X_l, X_r).$$

Then it follows that, if $\bar{G}_0(T \mid X) > 0$, F_X-a.e., then

$$\begin{aligned} \psi_0 &= \text{argmin}_\psi E_0 L(X, \psi) \\ &= \text{argmin}_\psi E_{P_0} \left\{ L(X, \psi) \frac{I(C > T)}{\bar{G}(T \mid X)} \right\} \\ &\equiv \text{argmin}_\psi E_{P_0} IC_0[O \mid G_0, L(\cdot, \psi)]. \end{aligned}$$

Thus, if we choose $L(O, \psi \mid \eta_0 = G_0) = IC_0(O \mid G_0, (\cdot, \psi))$, then $\psi_0 = \text{argmin}_\psi E_0 L(O, \psi \mid \eta_0)$. Alternatively, we can choose the corresponding double robust loss function:

$$L(O, \psi \mid \eta_0 = (Q_{X0}, G_0)) = IC(O \mid Q_{X0}, G_0, L(\cdot, \psi)),$$

as defined in (6).

Given candidate predictors $\hat{\psi}_k = \psi_k(\cdot \mid P_n)$, the corresponding distance is given by:

$$\begin{aligned} d_n(\hat{\psi}_k, \psi_0) &= \int L(O, \hat{\psi}_k \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\ &= \int \log \left(\frac{\psi_k(X_l, X_r \mid P_n)}{\psi_0(X_l, X_r)} \right) dF_{X0}(X). \end{aligned}$$

For the IPCW-choice of loss function we have:

$$\hat{k} = \text{argmax}_k E_{S_n} \sum_{i: S_n(i)=1} \log(\psi_k(X_{li}, X_{ri} \mid P^0_{n,S_n})) \frac{\Delta_i}{\bar{G}^0_{n,S_n}(T_i \mid X_i)}.$$

One crucial property of this cross-validation selector $\hat{k}$ is that it aims to minimize $d_n(\hat{\psi}_k, \psi_0)$ and thus it aims to choose the estimator which is closest to the true parameter value in Kullback-Leiber distance. For example, suppose that $\hat{\psi}_k$ are estimators based on different Cox-proportional hazards models or linear regression models. If one would apply standard model selection methodology such as AIC one would aim to choose the model which best fits the density of the data, while this cross-validation selector aims to choose the model which gives the best estimate of the parameter of interest.

A common choice of loss function $L(O, \psi)$ has been minus the logarithm of the F_{X0}-part of the density of the observed data. In this case, the corresponding cross-validation selector aims to estimate this F_{X0}-part of the density w.r.t. to a distance implied by the observed data density (which thus also involves the censoring distribution). The pro's and con's of this latter choice are discussed in more detail in Molinaro, Dudoit, van der Laan (2003). However, note that, if one is interested in a relatively marginal density of F_{X0}, then this type of likelihood cross-validation, though consistent, is not minimizing an appropriate distance.

This selection method is new and has not been handled in the current literature. In Molinaro, Dudoit, van der Laan (2003), it has been applied

to construct conditional histogram density estimators of survival based on recursive partitioning. Application of our general Theorem 1 yields a finite sample result and asymptotic equivalence with the optimal benchmark selector $\tilde{k}_n$. $\square$

4 Finite sample result and asymptotics: Quadratic loss function.

In this section, we will state a general theorem establishing a finite sample result and asymptotic equivalence of the cross-validation selector with a benchmark selector $\tilde{k}$ which for each given data of size $n(1-p)$ makes the optimal selection. In this theorem p can be chosen to be fixed, and does thus not necessarily converge to zero when the sample size converges to infinity. The theorem covers loss functions whose expectation can be estimated at a quadratic rate. This covers the loss functions presented in Examples 1-7. In a later section we will apply Theorem 1 to each of the examples and state the corresponding results.

Define Γ as the parameter space of the nuisance parameter η and

$$\Gamma_0 = \Gamma(P_0) = \{\eta : \text{argmin}_\psi E_0 L(O, \psi \mid \eta) = \psi_0\}$$

are the parameter values of η which still identify ψ_0 as the mimizer of the risk w.r.t. the loss function $L(O, \psi \mid \eta)$. We also define the function which $\hat{k}$ minimizes:

$$\hat{\theta}_{n(1-p)}(k) \equiv E_{S_n} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) dP^1_{n,S_n}(O).$$

A natural way to benchmark the selector $\hat{k}$ is to define for a $\eta_0 \in \Gamma(P_0)$, the supposed limit of the estimator η_n in the definition of $\hat{k}$, the following true conditional risk function

$$\tilde{\theta}_{n(1-p)}(k) = E_{S_n} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta_0) dP_0(O). \tag{7}$$

This quantaty represents the true conditional risk of the estimator $\hat{\psi}_k$ based on $n(1-p)$ observations. Note that $\hat{\theta}_{n(1-p)}(k)$ is an estimator of the true conditional risk $\tilde{\theta}_{n(1-p)}(k)$. Therefore, the minimizer

$$\tilde{k} \;=\; \text{argmin}_{k \in \{1,\cdots,K(n)\}} \tilde{\theta}_{n(1-p)}(k)$$

of the true conditional risk function for a given P_n defines a best possible choice for $\hat{k}$ since, given the data P_n, it indexes the best estimator among $\psi_k(. \mid P_{n(1-p)}), k \in \{1, \cdots, K(n)\}$ that achieves the optimal conditional risk based on $n(1-p)$ observations. In practice, we do not know the true conditional risk function $\tilde{\theta}_{n(1-p)}(\cdot)$ since it depends on the true observed data distribution P_0. Consequently, we do not have $\tilde{k}$ available to us. Note that $\tilde{k}$ distinguishes from the optimal benchmark selector $\tilde{k}_n$ (2) since it compares estimators based on $n(1-p)$ observations instead of n.

Let

$$\theta_{opt} \equiv E_0 L(O, \psi_0 \mid \eta_0)$$

be the optimal risk as achieved by the true parameter value ψ_0.

It is of interest and natural to establish how the performance of $\hat{k}$ in estimating the optimal risk compares with the performance of the minimizer $\tilde{k}$ of the true conditional risk. We derive a main result concerning the finite sample performance and the corresponding asymptotics (including asymptotic equivalence with $\tilde{k}$) of the cross-validated selector $\hat{k}$ defined above. Given a sequence of estimators $\hat{\psi}_k$, we define

$$\begin{aligned} d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) &= \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \\ &= E_{S_n} \int \{L(o, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(o, \psi_0 \mid \eta_0)\} dP_0(o). \end{aligned}$$

Similarly, we define $d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0) = \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}$. In this context, our main finite sample result compares the centered conditional risk $d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) \equiv \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}$ for the cross-validated selector with the centered conditional risk $d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0) = \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}$ for the benchmark selector. Finite sample bounds are obtained for the expected value of the predictive loss, $E\widetilde{PL}_{n(1-p)} = E(\tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k}))$. These imply, under appropriate conditions on the rate at which the nuisance parameter η_0 is estimated, convergence to zero in expectation and in probability of this risk difference $O(\log(K(n))/np)$ for loss functions whose risk can be estimated at a quadratic rate (Theorem 1). Consequently, if the risk difference $E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}$ converges to zero slower than these rates, then the ratio of expected risk differences $(E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to one. This implies, in particular, that $E\widetilde{PL}_{n(1-p)}/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to zero. The corresponding convergence in probability of the ratios of risk differences follows from Lemma 1 below.

In other words, we prove under appropriate conditions that for each fixed $p \in (0,1)$ $\hat{k}$ performs asymptotically as well as the benchmark selector $\tilde{k}$ in the sense that

$$\frac{d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} = \frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \longrightarrow 1 \quad \text{in probability as} \quad n \to \infty. \quad (8)$$

In a later section we will establish a general corollary of our Theorems which yields the asymptotic optimality $d_n(\hat{\psi}_{\hat{k}}, \psi_0)/d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0) \to 1$ if $p_n \to 0$ slowly enough so that (8) still holds at $p = p_n$.

In the next section we prove a similar Theorem 2 which applies to general loss functions, while Theorem 1 below applies to loss functions whose optimal risk θ_{opt} can be estimated at a quadratic rate. For example, in the special case of the squared error loss function in prediction, Theorem 1 provides a stronger convergence result than Theorem 2: for the squared error loss function, the rate of convergence is shown to be $O(\log(K(n))/np)$ rather than the slower $O(\log(K(n))/\sqrt{np})$ applicable to general loss functions. Both theorems consider general distributions of S_n, i.e., general cross-validation procedures with an arbitrary proportion p_n of observations included the validation sets. Finally, we note that our finite sample result and asymptotic result assume the setting in which $np_n \to \infty$; the later condition rules out LOOCV.

Formally, we propose the following definition of a quadratic loss function.

Definition 1 *If the following property of $L(O, \psi \mid \eta_0)$ holds at (ψ_0, η_0), then we refer to $L(O, \psi \mid \eta_0)$ as a quadratic loss function.*

Given any one-dimensional path $\epsilon \to \psi_\epsilon \in \mathbf{\Psi}$ through ψ_0 at $\epsilon = 0$ which is differentiable at $\epsilon = 0$,

$$\frac{d}{d\epsilon} d(\psi_\epsilon, \psi_0)\Big|_{\epsilon=0} = 0,$$

where $d(\psi, \psi_0) = \int L(O, \psi \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(o)$.

It is easy to verify that this property holds for each of the loss functions presented in our examples. By carrying out a Taylor expansion it can be argued that a quadratic loss function can be expected to satisfy Assumption A3 in Theorem 3.

Throughout the following theorem we introduce and use the following notation:

$$\begin{aligned} L^*(O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0) &= L(O,\psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O,\psi_0 \mid \eta_0) \\ L^{*0}_{n,S_n}(O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0) &= L(O,\psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) - L(O,\psi_0 \mid \eta^0_{n,S_n}) \\ (L^{*0}_{n,S_n} - L^*)(O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0) &= L^{*0}_{n,S_n}(O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0) - L^*(O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0). \end{aligned}$$

We also note that the rates $r_1(n), r_2(n)$ as defined in the theorem are determined by the rate at which the nuisance parameter estimate η_n approximates η_0.

Theorem 1 *Let $\psi_k(\cdot \mid P_n)$, $k = 1,\ldots,K(n)$, be a set of given estimators of $\psi_0 = argmin_{\psi\in\mathbf{\Psi}} \int L(O,\psi \mid \eta_0)dP_0(O)$. Suppose that $\psi_k(\cdot \mid P_n) \in \Psi$ for all k, with probability 1. Let $\hat{k} = argmin_k E_{S_n} \int L(O,\psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n})dP^1_{n,S_n}(O)$ be the cross-validation selector, and let $\tilde{k} = argmin_k E_{S_n} \int L(O,\psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n})dP_0(O)$ be the comparable benchmark selector. We also recall the notation $d_{n(1-p)}(\hat{\psi}_{\bar{k}},\psi_0) = \tilde{\theta}_{n(-p)}(\bar{k}) - \theta_{opt}$.*

Assumptions.

A1. *The limit η_0 of the estimator $\eta_n = \eta(P_n)$ for $n \to \infty$ is an element of $\Gamma(P_0)$.*

A2. *There exist a $M_1^* < \infty$ so that*

$$\sup_{\psi\in\mathbf{\Psi}} \sup_O L^*(O,\psi,\psi_0) \le M_1^*,$$

where the supremum over O is taken over a support of the distribution P_0 of O.

A3. *There exist a $M_2 < \infty$ so that for all $\psi \in \mathbf{\Psi}$*

$$VAR_{P_0}\left[L^*(O,\psi,\psi_0)\right] \le M_2 E_{P_0} L^*(O,\psi,\psi_0). \tag{9}$$

Definitions. *We define the following constants:*

$$\begin{aligned} M_1 &= 2M_1^* \\ c(M_1,M_2,\delta) &= 2(1+\delta)^2\left(\frac{M_1}{3} + \frac{M_2}{\delta}\right) \\ a_0 &\equiv 2M_1/3 \\ M_3(n) &= 3a_0 + \sqrt{2}\frac{\sqrt{\log(2)}}{\log(K(n))} + \frac{\sqrt{2}}{\sqrt{\log(K(n))}} + b_0 + \int_{b_0}^{\infty} 2K(n)^{1-m(x)}dx, \end{aligned}$$

where b_0 is the smallest constant larger than the solution of $1 - m(x) = 0$ with $m(x) \equiv 0.5\frac{x^2}{1/\log(K(n))+a_0 x}$. We note that $M_3(n) \downarrow$ in n. We also define the following sequences in n:

$$\begin{aligned} r_1(n) &\equiv \max_{\bar{k}\in\{\hat{k},\tilde{k}\}} \frac{E\int(L^{*0}_{n,S_n} - L^*)(O,\psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)}{\sqrt{E\int L^*(O,\psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)}} \\ r_2(n) &\equiv E\max_{k\in\{1,\ldots,K(n)\}} \sqrt{\int (L^{*0}_{n,S_n} - L^*)^2(O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)} \\ \tilde{r}(n) &\equiv \sqrt{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}},\psi_0)} \\ &= \sqrt{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}}. \end{aligned}$$

Finally, for any $\delta > 0$ we define

$$\begin{aligned} \epsilon_n(\delta) &\equiv (1+2\delta)\tilde{r}^2(n) + 2c(M_1,M_2,\delta)\frac{1+\log(K(n))}{np} + \\ &\quad (1+\delta)r_1(n)\tilde{r}(n) + \frac{2M_3(1+\delta)\log(K(n))}{(np)^{0.5}}\max(r_2(n),(np)^{-0.5}I(r_2(n)>0)). \end{aligned}$$

Finite Sample Result. *For any $\delta > 0$, we have*

$$\sqrt{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}},\psi_0)} \leq \frac{r_1(n)(1+\delta) + \sqrt{r_1(n)^2(1+\delta)^2 + 4\epsilon_n(\delta)}}{2}. \tag{10}$$

In the special case that η_0 is known so that $r_1(n) = r_2(n) = 0$, we have that the finite sample result (10) reduces to

$$Ed_{n(1-p)}(\hat{\psi}_{\hat{k}},\psi_0) = (1+2\delta)Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}},\psi_0) + 2c(M_1,M_2,\delta)\frac{1+\log(K(n))}{np}.$$

Asymptotic Implication. *For any $\delta > 0$*

$$Ed_{n(1-p)}(\hat{\psi}_{\hat{k}},\psi_0) \leq (1+2\delta)Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}},\psi_0) + O(H(n)),$$

where

$$\begin{aligned} H(n) &\equiv \max\left(\frac{\log(K(n))}{np}, \frac{\log(K(n))r_2(n)}{(np)^{0.5}}, r_1^2(n), r_1(n)\tilde{r}(n), r_1(n)^{1.5}\tilde{r}(n)^{0.5},\right. \\ &\quad \left.\frac{\sqrt{\log(K(n))}r_1(n)}{(np)^{0.5}}, \frac{\sqrt{\log(K(n)}r_2(n)^{0.5}r_1(n)}{(np)^{0.25}}\right). \end{aligned}$$

Consequently, we have the following scenarios.
Optimal rate: *If* $\max\left(\frac{\log(K(n))}{np}, r_1(n)^2, \log(K(n))r_2(n)^2\right) = O(\tilde{r}(n)^2)$, *then* $H(n) = O(\tilde{r}(n)^2)$, *and thus* $Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) = O(\tilde{r}(n)^2)$.
If either $\max\left(\frac{\log(K(n))}{np}, r_1(n)^2, \log(K(n))r_2(n)^2\right) = o(\tilde{r}(n)^2)$ *or* $\max\left(\frac{\log(K(n))^2}{np}, r_1(n)^2, r_2(n)^2\right) = o(\tilde{r}(n)^2)$, *then*

$$H(n) = o(\tilde{r}(n)^2).$$

In particular, we note that if $\max(r_1(n)^2, \log(K(n))r_2(n)^2) = o(\tilde{r}(n)^2)$, *then*

$$H(n) \;=\; O\left(\frac{\log(K(n))}{np}\right) + o(\tilde{r}^2(n)).$$

Asymptotic Optimality. *Consequently, under these two possible scenarios under which* $H(n) = o(\tilde{r}(n)^2)$, *we have*

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty. \tag{11}$$

Finally, if these two possible scenarios hold with $\tilde{r}(n)^2$ *replaced by the random quantaty* $d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)$, *then*

$$\frac{d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ in probability for } n \to \infty. \tag{12}$$

The final convergence in probability statement follows from Lemma 1.

Lemma 1 *Consider a sequence of random variables* $Z_1, Z_2, \ldots$, *with finite expectation* $E|Z_n| = O(g(n))$, *for a positive function* $g(n)$. *Then* $Z_n = O_P(g(n))$.

This result is a direct consequence of Markov's inequality. Next we present the proof of the Theorem.

4.1 Proof of Theorem 1.

By Assumption A1 and by definition of $\hat{k}$, we have

$$\begin{aligned} 0 \;&\leq\; d_{n(1-p)}(\psi_{\hat{k}}, \psi_0) \\ &=\; \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \end{aligned}$$

$$
\begin{aligned}
&= E_{S_n} \int L^*(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&= E_{S_n} \int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&\quad -E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&= E_{S_n} \int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&\quad +(1+\delta) E_{S_n} \int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
&\quad -(1+\delta) E_{S_n} \int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
&\quad -E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&\leq E_{S_n} \int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&\quad +(1+\delta) E_{S_n} \int L^{*0}_{n,S_n}(O, \psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
&\quad -(1+\delta) E_{S_n} \int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
&\quad -E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&= E_{S_n} \int L^*(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&\quad +(1+\delta) E_{S_n} \int L^*(O, \psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
&\quad -(1+\delta) E_{S_n} \int L^*(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
&\quad +(1+\delta) E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
&\quad -(1+\delta) E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
&= (1+2\delta) E_{S_n} \int L^*(O, \psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&\quad +T_{n,\hat{k}} + R_{n,\tilde{k}} + A_n,
\end{aligned}
$$

where

$$
\begin{aligned}
T_{n,k} &= -(1+\delta) E_{S_n} \int L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) d(P^1_{n,S_n} - P_0)(O) \\
&\quad -\delta E_{S_n} \int L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
R_{n,k} &= (1+\delta) E_{S_n} \int L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) d(P^1_{n,S_n} - P_0)(O)
\end{aligned}
$$

$$
\begin{aligned}
& -\delta E_{S_n} \int L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
A_n &= (1+\delta) E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O) \\
& -(1+\delta) E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP^1_{n,S_n}(O).
\end{aligned}
$$

We write

$$
A_n = A_{n1}(\tilde{k}) - A_{n1}(\hat{k}) + A_{n2}(\tilde{k}) - A_{n2}(\hat{k}),
$$

where

$$
\begin{aligned}
A_{n1}(k) &= (1+\delta) E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
A_{n2}(k) &= (1+\delta) E_{S_n} \int (L^{*0}_{n,S_n} - L^*)(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) d(P^1_{n,S_n} - P_0)(O).
\end{aligned}
$$

Analysis of $T_{n,\hat{k}}$, $R_{n,\tilde{k}}$: In Lemma 3 we prove for the specified constants M_1, M_2 and $c(M_1, M_2, \delta)$ that

$$
E(T_{n,\hat{k}} + R_{n,\tilde{k}}) \leq 2c(M_1, M_2, \delta) \frac{1 + \log(K(n))}{np}.
$$

Here we need Assumptions A2 and A3.
Analysis of $A_{n1}(\hat{k})$, $A_{n1}(\tilde{k})$: By definition of $r_1(n)$, we have for $\bar{k} \in \{\hat{k}, \tilde{k}\}$:

$$
E \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \leq r_1(n) \sqrt{E \tilde{\theta}_{n(1-p)}(\bar{k}) - \theta_{opt})}.
$$

Thus

$$
\begin{aligned}
E A_{n1}(\hat{k}) &\leq (1+\delta) r_1(n) \sqrt{E \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}} \\
E A_{n1}(\tilde{k}) &\leq (1+\delta) r_1(n) \sqrt{E \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}}.
\end{aligned}
$$

Analysis of $A_{n2}(\hat{k})$, $A_{n2}(\tilde{k})$: To bound $A_{n2}(\hat{k})$, we will apply Lemma 4, which is based on Bernstein's inequality. We first note that $A_{n2}(k) = (1+\delta) E_{S_n} \frac{1}{np} \sum_{i=1}^{np} Z_{k,n,i}$, where

$$
Z_{k,n,i} \equiv (L^{*0}_{n,S_n} - L^*)(O_i, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) - \int (L^{*0}_{n,S_n} - L^*)(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)
$$

and O_i, $i = 1, \ldots, np$ now constitutes the validation sample. Secondly, we notice that

$$\begin{aligned} T_n &\equiv \max_{k \in \{1,\ldots,K(n)\}} \mid \frac{1}{np} \sum_{i=1}^{np} Z_{k,n,i} \mid \\ &= \max_{k \in \{1,\ldots,K(n)\}} \mid \int (L^{*0}_{n,S_n} - L^*)(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) d(P^1_{n,S_n} - P_0)(O) \mid . \end{aligned}$$

Conditional on $B_n \equiv (S_n, P^0_{n,S_n})$, we have that $\frac{1}{np} \sum_{i=1}^{np} Z_{k,n,i}$ is a sum of i.i.d. mean zero copies of $Z_{k,n}$. By assumption A2, we have that $\mid Z_{k,n} \mid \leq 2M_1$. We also note that

$$\begin{aligned} \sigma_n(B_n) &\equiv \max_k \sqrt{Var(Z_{k,n} \mid B_n)} \\ &\leq \max_k \sqrt{\int (L^{*0}_{n,S_n} - L^*)^2 (O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)} \end{aligned}$$

By definition,

$$r_2(n) = E \max_k \sqrt{\int (L^{*0}_{n,S_n} - L^*)^2 (O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP(O))}$$

and thus $E\sigma_n(B_n) = r_2(n)$. Application of Lemma 4 proves that

$$\begin{aligned} ET_n &= E \max_{k \in \{1,\ldots,K(n)\}} \left| \int (L^{*0}_{n,S_n} - L^*)[O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) d(P^1_{n,S_n} - P)(O) \right| \\ &\leq M_3(n) \frac{\log(K(n))}{(np)^{0.5}} \max(r_2(n), (np)^{-0.5}), \end{aligned}$$

where $M_3(n)$ is decreasing in n. In particular, this implies

$$\begin{aligned} &E \mid \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) d(P^1_{n,S_n} - P)(O) \mid \\ &\leq M_3(n) \frac{\log(K(n))}{(np)^{0.5}} \max(r_2(n), (np)^{-0.5}) \end{aligned}$$

This proves that

$$EA_{n2}(\hat{k}) \leq (1 + \delta) M_3(n) \frac{\log(K(n))}{(np)^{0.5}} \max(r_2(n), (np)^{-0.5}).$$

Similarly,

$$EA_{n2}(\tilde{k}) \leq (1+\delta)M_3(n)\frac{\log(K(n))}{(np)^{0.5}}\max(r_2(n),(np)^{-0.5}).$$

Off course, if $r_2(n)=0$, then our bound is simply 0.
Establishing Finite Sample Result: Substituting these bounds in our final expression for $\tilde{\theta}_{\hat{k}}-\theta_{opt}$ yields the following inequality:

$$\begin{aligned} E\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt} &\leq (1+2\delta)\left(E\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}\right) \\ &\quad +2c(M_1,M_2,\delta)\frac{1+\log(K(n)}{np} \\ &+(1+\delta)r_1(n)\sqrt{E\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}}+(1+\delta)r_1(n)\sqrt{E\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}} \\ &\quad +\frac{2M_3(1+\delta)\log(K(n))}{(np)^{0.5}}\max(r_2(n),(np)^{-0.5}I(r_2(n)>0)) \quad (13) \end{aligned}$$

For notational convenience, we define

$$\hat{r}(n) = \sqrt{E\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}}.$$

Then (13) can be written as:

$$\hat{r}^2(n) \leq \epsilon_n(\delta)+(1+\delta)r_1(n)\hat{r}(n).$$

This inequality in $\hat{r}(n)$ is equivalent with

$$\hat{r}(n) \leq \frac{r_1(n)(1+\delta)+\sqrt{r_1(n)^2(1+\delta)^2+4\epsilon_n(\delta)}}{2}.$$

This proves the main statement (10) of the theorem.
Asymptotic Implications: From this it follows immediately that (use that $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for positive numbers a,b)

$$\hat{r}(n)=O\left(\max\left(r_1(n),\tilde{r}(n),\frac{\sqrt{\log(K(n))}}{(np)^{0.5}},r_1(n)^{0.5}\tilde{r}(n)^{0.5},\frac{\sqrt{\log(K(n))}}{(np)^{0.25}}r_2(n)^{0.5}\right)\right).$$

Substituting this in (13) yields:

$$\hat{r}(n)^2=E\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}\leq(1+2\delta)\left\{E\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}\right\}+O(H(n)).$$

Asymptotic Optimality: We will omit the algebra to show that the two scenarios indeed imply that $H(n) = o(\tilde{r}^2(n))$. If $H(n) = o(\tilde{r}^2(n))$, then for each $\delta > 0$

$$\frac{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \leq (1 + 2\delta) + o(1).$$

This proves

$$\frac{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 1 \text{ for } n \to \infty.$$

This completes the proof of the theorem. □

Lemmas.

We will now state and prove the required lemmas. Most of the lemmas are derived from Bernstein's inequality, which we state here as a lemma for ease of reference. A proof is given in Lemma A.2, p. 564 in Györfi et al. (2002).

Lemma 2 Bernstein's inequality. *Let Z_i, $i = 1, \ldots, n$, be independent real valued random variables such that $Z_i \in [a, b]$ with probability one. Let $0 < \sum_{i=1}^{n} VAR(Z_i)/n \leq \sigma^2$. Then, for all $\epsilon > 0$,*

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - EZ_i) > \epsilon\right) \leq \exp\left(-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3}\right).$$

This implies

$$Pr\left(\frac{1}{n} \mid \sum_{i=1}^{n}(Z_i - EZ_i) \mid > \epsilon\right) \leq 2\exp\left(-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3}\right).$$

Lemma 3 *Assume Assumptions A2 and A3. Let $M_1 = 2M_1^*$. Then*

$$ET_{n,\hat{k}} \leq c(M_1, M_2, \delta)\frac{1 + \log(K(n))}{np}, \tag{14}$$

where

$$c(M_1, M_2, \delta) = 2(1 + \delta)^2\left(\frac{M_1}{3} + \frac{M_2}{\delta}\right).$$

Similarly,

$$ER_{n,\tilde{k}} \leq c(M_1, M_2, \delta)\frac{1 + \log(K(n))}{np}.$$

Thus

$$E(T_{n,\hat{k}} + R_{n,\tilde{k}}) \leq 2c(M_1, M_2, \delta)\frac{1 + \log(K(n))}{np}.$$

Proof. Let $Z_k \equiv L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)$. Given S_n, P^0_{n,S_n}, the conditions imply that

$$\begin{aligned} \mid Z_k - E(Z_k \mid S_n, P^0_{n,S_n}) \mid &\leq M_1 \\ \mathrm{VAR}(Z_k) &\leq M_2 E Z_k. \end{aligned}$$

The result (14) is proved in theorem 2 in Dudoit, van der Laan (2003). To provide the reader with some background, we will mention here some of the main ideas in the proof. Firstly, one notes that $T_{n,\hat{k}} = E_{S_n} T_{n,\hat{k}}(S_n)$, where, conditional on (S_n, P^0_{n,S_n}), $T_{n,k}(S_n)$ equals $1 + \delta$ times an empirical mean of np copies of $Z_k - EZ_k$ minus δ times EZ_k. Secondly, one applies Bonferoni to $P(T_{n,\hat{k}}(S_n) \geq s \mid S_n, P^0_{n,S_n})$, and one applies Bernstein's inequality to the k-specific conditional tail probability of $T_{n,k}(S_n)$. Thirdly, by exploiting the fact that $\mathrm{VAR}(Z_k) \leq M_2 EZ_k$ it can be shown that the conditional tail probability can be bounded by $\exp(-cnps)$ for some $c < \infty$, instead of the usual $\exp(-c(np)s^2)$. Finally, bounding the expectation of $T_{n,\hat{k}}$ in terms of the integral over the obtained tail probability for $T_{n,\hat{k}}(S_n)$ yields the wished result. □

The analysis of the A_{n2} terms relies on the following general lemma, which fully exploits Bernstein's inequality.

Lemma 4 *Suppose that for each integer n and each $k \in \{1, \ldots, K(n)\}$, conditional on a random variable B_n, $Z_{k,n,i}$, $i = 1, \ldots, n$, are n i.i.d. copies of $Z_{k,n}$ with $\mid Z_{k,n} \mid \leq W_n(B_n)$ and variance $\sigma^2_{kn}(B_n) = VAR(Z_{k,n} \mid B_n)$. Let $\sigma_n(B_n) \equiv \max_{k \in \{1,\ldots,K(n)\}} \sigma_{kn}(B_n)$. Let*

$$T_n = \max_{k \in \{1,\ldots,K(n)\}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_{k,n,i} \right|.$$

If $W_n(B_n) \leq W < \infty$ for some constant W, then there exists an $C < \infty$ so that

$$ET_n \leq C \frac{\log(K(n))}{n^{0.5}} E\sigma^*_n(B_n), \tag{15}$$

where $\sigma^*_n(B_n) = \max(\sigma_n(B_n), n^{-0.5})$.

One can set $C = M_3(n)$*, where*

$$M_3(n) \equiv 3a_0 + \sqrt{2}\frac{\sqrt{\log(2)}}{\log(K(n))} + \frac{\sqrt{2}}{\sqrt{\log(K(n))}} + b_0 + \int_{b_0}^{\infty} 2K(n)^{1-m(x)}dx,$$

where $a_0 \equiv W/3$ *and* b_0 *is the smallest constant larger than the solution of* $1 - m(x) = 0$ *for* $m(x) = 0.5\frac{x^2}{1/\log(K(n))+a_0 x}$*. Note that* $M_3(n)$ *is decreasing in* n*.*

Proof. The following Lemma 5 proves that

$$Pr\left(\frac{n^{0.5}}{\sigma_n^*(B_n)\log(K(n))}T_n \geq x \mid B_n\right) \leq 2K(n)\exp\left(-\frac{1}{2}\frac{x^2\log(K(n))}{1/\log(K(n)) + \frac{W_n(B_n)}{3\sigma_n^*(B_n)n^{0.5}}x}\right).$$

Thus for any $u > 0$

$$\frac{n^{0.5}}{\sigma_n^*(B_n)\log(K(n))}E(T_n \mid B_n) \leq u + \int_u^{\infty} 2K(n)\exp\left(-\frac{1}{2}\frac{x^2\log(K(n))}{1/\log(K(n)) + \frac{W_n(B_n)}{3\sigma_n^*(B_n)n^{0.5}}x}\right)dx. \quad (16)$$

The right-hand side attains its minimum at $u = U_n(B_n)$, where

$$\begin{aligned} U_n(B_n) &= \frac{W_n(B_n)}{3\sigma_n^*(B_n)n^{0.5}}\frac{\log(2K(n))}{\log(K(n))} \\ &\quad + \frac{1}{\log(K(n))}\sqrt{\log^2(2K(n))\frac{W_n^2(B_n)}{(3\sigma_n^*(B_n)n^{0.5})^2} + 2\log(2K(n))}. \end{aligned}$$

The integrand in (16) equals 1 at $x = U_n(B_n)$ and is decreasing for $x > U_n(B_n)$. Thus this shows

$$\frac{n^{0.5}}{\sigma_n^*(B_n)\log(K(n))}E(T_n \mid B_n) \leq U_n(B_n) + C_n(B_n),$$

where

$$C_n(B_n) = \int_{U_n(B_n)}^{\infty} 2K(n)\exp\left(-\frac{1}{2}\frac{x^2\log(K(n))}{1/\log(K(n)) + \frac{W_n(B_n)}{3\sigma_n^*(B_n)n^{0.5}}x}\right)dx.$$

Consider this bound when $W_n(B_n) \leq W$ a.s., so that $W_n(B_n)/(3\sigma_n^*(B_n)n^{0.5}) \leq a_0 \equiv W/3$. Then

$$\begin{aligned} U_n(B_n) &\leq a_0 \frac{\log(2K(n))}{\log(K(n))} + \frac{1}{\log(K(n))}\sqrt{\log^2(2K(n))a_0^2 + 2\log(2K(n))} \\ &\leq 4a_0 + \sqrt{2}\frac{\sqrt{\log(2)}}{\log(K(n))} + \frac{\sqrt{2}}{\sqrt{\log(K(n))}}, \end{aligned}$$

which is decreasing in $K(n) \geq 2$. Here we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a \geq 0, b \geq 0$ and $K(n) \geq 2$. In addition, since the integrand in (16) is smaller than 1 on $[U_n(B_n), \infty)$, we have for any $C < \infty$ that the integral $\int_{U_n(B_n)}^{\infty}$ in (16) can be bounded as follows:

$$\begin{aligned} C_n(B_n) &= \int_{U_n(B_n)}^{\infty} 2K(n) \exp\left(-\frac{1}{2}\frac{x^2\log(K(n))}{1/\log(K(n)) + \frac{W_n(B_n)}{3\sigma_n^*(B_n)n^{0.5}}x}\right) dx \\ &\leq I(U_n(B_n) \leq C)C + \int_C^{\infty} 2K(n) \exp\left(-\frac{1}{2}\frac{x^2\log(K(n))}{1/\log(K(n)) + a_0 x}\right) dx \\ &\leq I(U_n(B_n) \leq C)C + \int_C^{\infty} 2K(n)^{1-m(x)} dx, \end{aligned}$$

with $m(x) = 0.5\frac{x^2}{1/\log(K(n))+a_0x}$. Let $C = b_0$ be chosen to be larger than the solution of $1 - m(x) = 0$ so that $1 - m(x) < 0$. Then the latter bound is decreasing in $K(n)$, and for each fixed $K(n)$ the integral is finite. We conclude that, if we set

$$M_3(n) = 3a_0 + \sqrt{2}\frac{\sqrt{\log(2)}}{\log(K(n))} + \frac{\sqrt{2}}{\sqrt{\log(K(n))}} + b_0 + \int_{b_0}^{\infty} 2K(n_0)^{1-m(x)} dx,$$

then $E(T_n \mid B_n) \leq M_3(n)\frac{\log(K(n))}{n^{0.5}}\sigma_n^*(B_n)$ and hence $ET_n \leq M_3(n)\frac{\log(K(n))}{n^{0.5}}E\sigma_n^*(B_n)$. This proves (15). □

Lemma 5 *For each n and $k \in \{1, \ldots, K(n)\}$, let $Z_{k,n,i}$, $i = 1, \ldots, n$, be n independent mean zero random variables with variance $VAR(Z_{k,n,i}) \leq \sigma_n^2$ and $Pr(\max_{i,k} \mid Z_{k,n,i} \mid \leq W_n) = 1$ for $W_n < \infty$. If $W_n < W < \infty$, then*

$$\max_{k\in\{1,\ldots,K(n)\}} \left|\frac{1}{n}\sum_{i=1}^{n} Z_{k,n,i}\right| = O_P\left(\frac{\log(K(n))}{n^{0.5}\sigma_n^*}\right),$$

where $\sigma_n^* = \max(\sigma_n, n^{-0.5})$. *Specifically,*

$$
\begin{aligned}
&Pr\left(\frac{n^{0.5}}{\sigma_n^* \log(K(n))} \max_{k\in\{1,\ldots,K(n)\}} \left|\frac{1}{n}\sum_{i=1}^{n} Z_{k,n,i}\right| \geq x\right) \\
&\leq 2K(n)\exp\left(-\frac{1}{2}\frac{x^2\log(K(n))}{\frac{1}{\log(K(n))}+\frac{W_n x}{3\sigma_n^* n^{0.5}}}\right). \\
&= 2K(n)\max_k\left(\frac{1}{K(n)}\right)^{\left[\frac{\frac{1}{2}x^2}{1/\log(K(n))+\frac{W_n x}{3\sigma_n^* n^{0.5}}}\right]}.
\end{aligned}
$$

Proof. By the Bonferoni and Bernstein inequalities we have

$$
\begin{aligned}
&Pr\left(\frac{1}{\sigma_n^* n^{0.5}\log(K(n))}\max_k \mid \sum_{i=1}^{n} Z_{k,n,i} \mid > x\right) \\
&\leq K(n)\max_k Pr\left(\mid \textstyle\sum_{i=1}^{n} Z_{k,n,i} \mid > x\sigma_n^* n^{0.5}\log(K(n))\right) \\
&\leq 2K(n)\max_k \exp\left(-\frac{1}{2}\frac{x^2\sigma_n^{*2} n\log^2(K(n))}{n\sigma_n^{*2}+\frac{W_n\sigma_n^* n^{0.5}\log(K(n))x}{3}}\right) \\
&= 2K(n)\max_k \exp\left(-\frac{1}{2}\frac{x^2\log(K(n))}{\frac{1}{\log(K(n))}+\frac{W_n x}{3\sigma_n^* n^{0.5}}}\right).\square
\end{aligned}
$$

5 Finite sample result and asymptotics: general loss functions.

In this subsection we prove the most general theorem which can be applied to any loss function. This theorem should be applied in case Assumption A3 of Theorem 1 fails to hold, or, as we suggested, if the loss functions fails to satisfy Definition 1.

Theorem 2 *Let* $\psi_k(\cdot \mid P_n)$, $k = 1, \ldots, K(n)$, *be a set of given estimators of* $\psi_0 = argmin_{\psi\in\Psi}\int L(O,\psi \mid \eta_0)dP_0(O)$. *Suppose that* $\psi_k(\cdot \mid P_n) \in \Psi$ *for all* k, *with probability 1. Let* $\hat{k} = argmin_k E_{S_n}\int L(O,\psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n})dP^1_{n,S_n}(O)$ *be the cross-validation selector, and let* $\tilde{k} = argmin_k E_{S_n}\int L(O,\psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n})dP_0(O)$ *be the comparable benchmark selector. We also recall the notation* $d_{n(1-p)}(\hat{\psi}_{\bar{k}},\psi_0) = \tilde{\theta}_{n(-p)}(\bar{k}) - \theta_{opt}$, *where* $\bar{k}$ *denotes a possibly random* k.

Assumptions.
A1. *The limit η_0 of the estimator $\eta_n = \eta(P_n)$ for $n \to \infty$ is an element of $\Gamma(P_0)$.*
A2. *There exist a $M_1^* < \infty$ so that*

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{O} L^*(O, \psi, \psi_0) \leq M_1^*,$$

where the supremum over O is taken over a support of the distribution P_0 of O.

Definitions. *We define the following sequences in n:*

$$\begin{aligned} f(M_1^*, K(n), np) &\equiv 4M_1^{*2}\sqrt{\frac{\log K(n)}{np}} \\ r_1(n) &\equiv \max_{\bar{k} \in \{\hat{k}, \tilde{k}\}} \frac{E\int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}{\sqrt{E\int L^*(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}} \\ \tilde{r}(n) &\equiv \sqrt{E d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \end{aligned}$$

We also define

$$\epsilon_n \equiv \tilde{r}^2(n) + f(M_1^*, K(n), np) + r_1(n)\tilde{r}(n).$$

Finite Sample Result. *We have*

$$\sqrt{E d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)} \leq \frac{r_1(n) + \sqrt{r_1(n)^2 + 4\epsilon_n}}{2}. \tag{17}$$

If η_0 is known so that $r_1(n) = 0$, then (17) reduces to

$$E d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) \leq E d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0) + f(M_1^*, K(n), np).$$

Asymptotic Implication. *We have*

$$E d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) \quad \leq \quad E d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0) + O(H(n)),$$

where

$$H(n) \equiv \max\left(\frac{\log^{0.5}(K(n))}{(np)^{0.5}}, r_1(n)\tilde{r}(n), r_1(n)^2, r_1(n)\frac{\log^{0.25}(K(n))}{(np)^{0.25}}, r_1(n)^{1.5}\tilde{r}(n)^{0.5} \right).$$

Optimal rate: *If* $\max\left(r_1(n)^2, \frac{\log(K(n))}{\sqrt{np}}\right) = O(\tilde{r}(n)^2)$, *then* $H(n) = O(\tilde{r}(n)^2)$.
Asymptotic Optimality/Equivalence. *If* $\frac{r_1(n)}{\tilde{r}(n)} \to 0$ *for* $n \to \infty$, *then*

$$\begin{aligned} O(H(n)) &= O\left(\frac{\log^{0.5}(K(n))}{(np)^{0.5}}\right) + o\left(\tilde{r}(n)\max\left(\tilde{r}(n), \frac{\log^{0.25}(K(n))}{(np)^{0.25}}\right)\right) \\ &= O\left(\frac{\log^{0.5}(K(n))}{(np)^{0.5}}\right) + o(\tilde{r}^2(n)). \end{aligned}$$

Thus, if also $\frac{\log^{0.5}(K(n))}{(np)^{0.5}\tilde{r}(n)^2} \to 0$ *for* $n \to \infty$, *then*

$$H(n) = o(\tilde{r}(n)^2) = o(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}).$$

Consequently, if $\frac{r_1(n)}{\tilde{r}(n)} \to 0$ *and* $\frac{\log^{0.5}(K(n))}{(np)^{0.5}\tilde{r}(n)^2} \to 0$ *for* $n \to \infty$, *then*

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty. \tag{18}$$

Finally, if $\frac{r_1(n)}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 0$ *in probability, and* $\frac{\log^{0.5}(K(n)}{(np)^{0.5}\{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\}} \to 0$ *in probability, then*

$$\frac{d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ in probability for } n \to \infty. \tag{19}$$

Proof of Theorem. By Assumption A1, and by definition of $\hat{k}$, we have

$$\begin{aligned} 0 &\leq \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \\ &= E_{S_n}\int L^*(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)dP_0(O) \\ &= E_{S_n}\int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)dP_0(O) \\ &\quad -E_{S_n}\int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)dP_0(O) \\ &= E_{S_n}\int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)dP^1_{n,S_n}(O) \\ &\quad +E_{S_n}\int L^{*0}_{n,S_n}(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)d(P_0 - P^1_{n,S_n})(O) \\ &\quad -E_{S_n}\int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)dP_0(O) \end{aligned}$$

$$
\begin{aligned}
&\leq E_{S_n}\int L^{*0}_{n,S_n}(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP^1_{n,S_n}(O)\\
&\quad +E_{S_n}\int L^{*0}_{n,S_n}(O,\psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)d(P_0-P^1_{n,S_n}(O)\\
&\quad -E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)\\
&= E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP^1_{n,S_n}(O)\\
&\quad +E_{S_n}\int L^*(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP^1_{n,S_n}(O)\\
&\quad +E_{S_n}\int L^{*0}_{n,S_n}(O,\psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)d(P_0-P^1_{n,S_n}(O)\\
&\quad -E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)\\
&= E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)d(P^1_{n,S_n}-P_0)(O)\\
&\quad +E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)\\
&\quad +E_{S_n}\int L^*(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)\\
&\quad +E_{S_n}\int L^*(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)d(P^1_{n,S_n}-P_0)(O)\\
&\quad +E_{S_n}\int L^{*0}_{n,S_n}(O,\psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)d(P_0-P^1_{n,S_n}(O)\\
&\quad -E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)\\
&= E_{S_n}\int L^*(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)\\
&\quad +E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)\\
&\quad +E_{S_n}\int L^{*0}_{n,S_n}(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)d(P^1_{n,S_n}-P_0)(O)\\
&\quad +E_{S_n}\int L^{*0}_{n,S_n}(O,\psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)d(P_0-P^1_{n,S_n}(O)\\
&\quad -E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)\\
&= E_{S_n}\int L^*(O,\psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O)+T_{n,\tilde{k}}-T_{n,\hat{k}}+A_{n,\tilde{k}}-A_{n,\hat{k}},
\end{aligned}
$$

where

$$
T_{n,k} = E_{S_n}\int L^{*0}_{n,S_n}[O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)d(P^1_{n,S_n}-P_0)(O)
$$

$$A_{n,k} \;=\; E_{S_n}\int (L^{*0}_{n,S_n}-L^*)(O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O).$$

Analyzing $T_{n,\hat{k}}$, $T_{n,\tilde{k}}$. By Lemma 6 below we have

$$E\left\{T_{n,\tilde{k}}-T_{n,\hat{k}}\right\}\leq f(M_1^*,K(n),np).$$

Here we need Assumption A2.

Lemma 6 *(van der Vaart, 2003, van der Laan, Dudoit, van der Vaart, 2003) Suppose that* $\sup_O \mid L^{*0}_{n,S_n}(O,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0) \mid< M_1^*$ *for all* k *a.s., where the supremum is over a support of the distribution* P_0 *of* O*. Let*

$$f(M_1^*,K(n),np)\equiv 4M_1^{*2}\sqrt{\frac{\log K(n)}{np}}.. \tag{20}$$

We have the following finite sample result

$$ET_{n,\hat{k}}\leq f(M_1^*,K(n),np)/2.$$

We also have

$$ET_{n,\tilde{k}}\leq f(M_1^*,K(n),np)/2.$$

Thus, in particular,

$$ET_{n,\hat{k}}-ET_{n,\tilde{k}}\leq f(M_1^*,K(n),np).$$

This proof is given in van der Vaart (2003) and van der Laan et al. (2003), which is an improvement (by a factor $\sqrt{\log K(n)}$) of the original result in Theorem 1 of Dudoit, van der Laan (2003).

Analysis of $A_n(\hat{k})$, $A_n(\tilde{k})$: By definition of $r_1(n)$, we have for $\bar{k}\in\{\hat{k},\tilde{k}\}$

$$E\int (L^{*0}_{n,S_n}-L^*)(O,\psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)dP_0(O) \;\leq\; r_1(n)\sqrt{E\tilde{\theta}_{n(1-p)}(\bar{k})-\theta_{opt})}.$$

Thus

$$\begin{aligned} EA_n(\hat{k}) &\leq r_1(n)\sqrt{E\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}} \\ EA_n(\tilde{k}) &\leq r_1(n)\sqrt{E\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}}. \end{aligned}$$

Establishing Finite Sample Result: Substituting these bound in our final bound for $\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}$ yields the following inequality.

$$\begin{aligned} E\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt} \;\leq\; &\left(E\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}\right)+f(M_1^*,K(n),np) \\ &+r_1(n)\sqrt{E\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt}}+r_1(n)\sqrt{E\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt}}. \end{aligned} \tag{21}$$

For notational convenience, let

$$\hat{r}(n) \;=\; \sqrt{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}.$$

Let

$$\epsilon_n \equiv \tilde{r}^2(n) + f(M_1^*, K(n), np) + r_1(n)\tilde{r}(n).$$

Then (21) can be written as:

$$\hat{r}^2(n) \leq \epsilon_n + r_1(n)\hat{r}(n).$$

This inequality in $\hat{r}(n)$ is equivalent with

$$\hat{r}(n) \leq \frac{r_1(n) + \sqrt{r_1(n)^2 + 4\epsilon_n}}{2}.$$

This proves the main statement (17) of the theorem.
Asymptotic Implications: From this it follows immediately that (use that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for positive numbers a, b)

$$\hat{r}(n) = O\left(\max\left(r_1(n), \tilde{r}(n), \frac{\log^{0.25}(K(n))}{(np)^{0.25}}, r_1(n)^{0.5}\tilde{r}(n)^{0.5}\right)\right).$$

Substituting this in (21) yields:

$$\hat{r}(n)^2 = E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq \left\{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\right\} + O(H(n)).$$

This proves the asymptotic implication statement of the theorem.
Asymptotic Optimality: If $\frac{r_1(n)}{\tilde{r}(n)} \to 0$ and $\frac{\log^{0.5}(K(n))}{(np)^{0.5}\tilde{r}(n)} \to 0$ for $n \to \infty$, then

$$H(n) = o(\tilde{r}(n)) = o(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}).$$

This completes the proof of the theorem. □

6 Asymptotic equivalence with oracle procedure.

Theorems 1 and 2 provide a finite sample bound for the expected value of $\tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k})$, which compares the performance of the cross-validated

selector $\hat{k}$ to the benchmark $\tilde{k}$ in terms of the conditional risks $\tilde{\theta}_{n(1-p)}(k)$ based on $n(1-p)$ training observations. This bound is used to prove that the ratio $(E\tilde{\theta}_{n(1-p)}(\hat{k})-\theta_{opt})/(E\tilde{\theta}_{n(1-p)}(\tilde{k})-\theta_{opt})$ converges to one, or equivalently that $E\tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to zero.

However, one would like the cross-validated selector $\hat{k}$ to perform as well as the benchmark selector $\tilde{k}_n$ (2) based on the whole sample of size n, rather than only $n(1-p)$ as above. The following is an immediate corollary of Theorem 1 and Theorem 2, which relates $\tilde{\theta}_{n(1-p)}(\hat{k})$ to the risk of a benchmark selector based on n observations, $\tilde{\theta}_n(\tilde{k}_n)$. In this corollary, we use the notation $p = p_n$ to emphasize the dependence of the validation set proportion p on n. It proves that, if $p = p_n$ converges slowly enough to zero when the sample size n converges to infinity, then, given a mild condition (22) below, the wished asymptotic optimality of the selector $\hat{k}$ follows. The proof of this corollary is straightforward and provided in Dudoit, van der Laan (2003) and van der Laan, Dudoit, Keles (2003).

Corollary 1 *Let $\tilde{k}_{n(1-p)} \equiv \tilde{k}$. If $p = p_n \to 0$, and the conditions of theorem 1 or Theorem 2 hold so that*

$$\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k}_{n(1-p)}) - \theta_{opt}} \to 1 \text{ in probability, for } n \to \infty,$$

and for $n \to \infty$

$$\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}} \quad \to \quad 1 \quad \text{in probability,} \tag{22}$$

then

$$\left(\frac{d_{n(1-p_n)}(\hat{\psi}_{\hat{k}}, \psi_0)}{d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)} = \right) \frac{\tilde{\theta}_{n(1-p_n)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}} \quad \to \quad 1 \quad \text{in probability.} \tag{23}$$

A sufficient condition for (22) to hold is that

$$\left(n^\gamma \left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}\right), (n(1-p_n))^\gamma \left(\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}\right)\right) \stackrel{D}{\Rightarrow} (Z, Z)$$

for some $\gamma > 0$ and random variable Z with $Pr(Z > a) = 1$ for some $a > 0$. In particular, if $Pr(S_n = s) = 1$ for some $s \in \{0,1\}^n$ (i.e., single split cross-validation), then it suffices to assume $n^\gamma \left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}\right) \stackrel{D}{\Rightarrow} Z$ for some $\gamma > 0$ and $Pr(Z > a) = 1$ for some $a > 0$.

We have a similar result for convergence in expectation: If $p = p_n \to 0$, the conditions of theorem 1 hold so that for $n \to \infty$

$$\frac{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_{n(1-p)}(\tilde{k}_{n(1-p)}) - \theta_{opt}} \to 1,$$

and for $n \to \infty$

$$\frac{E\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{E\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}} \quad \to \quad 1 \tag{24}$$

then

$$\left(\frac{Ed_{n(1-p_n)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)} = \right) \frac{E\tilde{\theta}_{n(1-p_n)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}} \quad \to \quad 1. \tag{25}$$

Consider the estimator $\psi(\cdot \mid P_n) = \psi_{\hat{k}(P_n)}(\cdot \mid P_n)$. Suppose that for n large enough

$$E \int L(o, \psi(\cdot \mid P^0_{n,S_n}) \mid \eta_0) dP_0(o) \geq E \int L(o, \psi(\cdot \mid P_n) \mid \eta_0) dP_0(o),$$

then (25) implies the wished asymptotic equivalence result:

$$\left(\frac{Ed_n(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)} = \right) \frac{E\tilde{\theta}_n(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}} \to 1.$$

In other words, if the estimator $\psi(\cdot \mid P_n)$ is capable of learning, then (25) implies the wished optimality result. We also note that the condition (24) is not more than a very weak regularity condition.

7 Application of theorem to the examples.

In this section we apply Theorem 1 to our examples. This results in seven corollaries, where each proof involves verification of Assumptions A1,A2 and A3 of Theorem 1. In the case that the loss function depends on an unknown nuisance parameter η_0, we will also establish a worked out bound for $r_1(n)$ and $r_2(n)$. Subsequent application of corollary 1 establishes the asymptotic equivalence (25) with the oracle selection procedure $\tilde{k}_n$.

Corollary 2 Predictor Selection.
Setting: *Let* $O = (Y, W) \sim P_0$, $\psi_0(W) = E_0(Y \mid W)$ *and* $L(O, \psi) = (Y - \psi(W))^2$. *We have* $d_{n(1-p)}(\hat{\psi}_k, \psi_0) = E_{S_n} \int (\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W))^2 dF_0(W)$.
Assumptions: *Assume that there exists a* $C_0 < \infty$ *so that* $\mid Y \mid \leq C_0$ *and* $\sup_n \mid \psi_k(\cdot \mid P_n) \mid \leq C_0$ *with probability 1.*
Definitions: *Let* $M_1 = 8C_0^2$ *and* $M_2 = 16C_0^2$.
Results: *The statements of Theorem 1 holds with these specified constants* M_1, M_2 *and* $r_1(n) = r_2(n) = 0$. *Thus, we have for any* $\delta > 0$

$$Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) \leq (1+2\delta) \left\{ Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0) \right\} + 2c(M_1, M_2, \delta) \frac{1 + \log(K(n))}{np}.$$

In particular, if $\frac{\log(K(n))}{np\{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)\}} \to 0$, *then*

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty.$$

Proof. We want to apply Theorem 1, which requires verification of assumptions A1-A3. Since the loss function $L(\cdot, \psi)$ has no nuisance parameter η_0, condition A1 holds automatically. Regarding condition A2, we note that

$$L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) = (Y - \psi_k(W \mid P^0_{n,S_n}))^2 - (Y - \psi_0(W))^2.$$

Thus A2 holds with $M_1^* = 4C_0^2$. Regarding condition A3 we note that

$$\begin{aligned}
&\int \left\{ (Y - \psi_k(W \mid P^0_{n,S_n}))^2 - (Y - \psi_0(W))^2 \right\}^2 dP_0(O) \\
&= \int (\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W))^2 (2Y - \psi_k(\cdot \mid P^0_{n,S_n}) - \psi_0(W))^2 dP_0(O) \\
&\leq 16C_0^2 \int (\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W))^2 dP_0(O) \\
&= 16C_0^2 \int (Y - \psi_k(W \mid P^0_{n,S_n})^2 - (Y - \psi_0(W))^2 dP_0(O).
\end{aligned}$$

Thus A3 holds with $M_2 = 16C_0^2$. So we can apply Theorem 1 with $r_1(n) = r_2(n) = 0$ and the specified constants M_1, M_2. This proves the corollary. □

Corollary 3 Density Estimator Selection.
Setting: *Let* $O \sim P_0$, $\psi_0(W) = f_0 \equiv \frac{dP_0}{d\mu}$ *be the density of* P_0 *w.r.t. a dominating measure* μ, *and let* $L(O, \psi) = -\log(\psi(O))$. *We have* $d_{n(1-p)}(\hat{\psi}_k, \psi_0) = E_{S_n} \int \log \left(\frac{\psi_k(o \mid P^0_{n,S_n})}{\psi_0(o)} \right) dP_0(o)$.
Assumptions: *Assume that there exists a* $\delta > 0$ *and* $C_0 < \infty$ *so that*

$\delta < f_0(O) \leq C_0$ *and* $\delta <\mid \psi_k(O \mid P_n) \mid \leq C_0$*, for all* n*, with probability 1, for* P_0*-almost every* O*.*
Definitions: *Let* $M_1 = 2\log(C_0/\delta)$*, and let* $M_2 = 4C_0/\delta$*.*
Results: *The statements of Theorem 1 hold with these values of* M_1, M_2 *and* $r_1(n) = r_2(n) = 0$*. Thus, for any* $\delta > 0$

$$Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) \leq (1+2\delta)Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0) + 2c(M_1, M_2, \delta)\frac{1+\log(K(n))}{np}.$$

In particular, if $\frac{\log(K(n))}{npEd_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 0$*, then*

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty.$$

Proof. We want to apply Theorem 3, which requires verification of assumptions A1-A3. Since the loss function $L(O, \psi)$ has no nuisance parameter η_0, condition A1 holds automatically. Regarding condition A2, we note that

$$L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) = -\log\left(\frac{\psi_k(O \mid P^0_{n,S_n})}{\psi_0(O)}\right).$$

Thus A2 holds with $M_1^* = \log(C_0/\delta)$. In van der Laan, Dudoit, Sunduz (2003) (Lemma 2, page 9) it is shown that condition A3 holds with $M_2 = 4C_0/\delta$. So we can apply Theorem 3 with $r_1(n) = r_2(n) = 0$ and the specified constants M_1, M_2. This proves the corollary. □

Corollary 4 *(***Survival Predictor Selection***)*
Setting: *Let* $R(t) = I(T \leq t)$ *be the indicator process for a survival time* T*, and* $L(t)$ *is a covariate process,* $t \geq 0$*. Let* $X = \bar{X}(T) = (\bar{R}(T), \bar{L}(T)) \sim F_{X,0}$ *be the full data structure of interest, and let* $Y \equiv \log(T)$ *be the log survival time, and* $W = L(0)$ *be the vector of baseline covariates. Let* C *be a right-censoring time of* T *with conditional distribution* $G_0(\cdot \mid X)$*, given* X*. We assume that this conditional distribution satisfies coarsening at random: for* $t \leq T$

$$\lambda_C(t \mid X) = m(t, \bar{X}(t)) \text{ for a measurable function } m.$$

Let

$$O = (\tilde{T} = \min(T, C), \Delta = I(\tilde{T} = T), \bar{X}(\tilde{T})) \sim P_0 = P_{F_{X,0}, G_0}.$$

Let $\psi_0(W) = E_0(Y \mid W)$ be the parameter of interest. Define

$$IC(O \mid G, D) = D(X)\frac{\Delta}{\bar{G}(T \mid X)}$$

as the so called inverse probability of censoring weighted mapping from full data functions $D(X)$ to observed data functions. We define

$$L(O, \psi \mid \eta_0 = G_0) = IC(O \mid G_0, L(\cdot, \psi)),$$

where $L(X, \psi) = (Y - \psi(W))^2$. Then $\psi_0 = argmin_\psi E_0 L(O, \psi \mid \eta_0)$, if $\bar{G}_0(Y \mid X) > 0$, F_{X0}-a.e. We have $d_{n(1-p)}(\hat{\psi}_k, \psi_0) = E_{S_n} \int \left(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W)\right)^2 dF_0(W)$.
Assumptions: *Suppose that $\bar{G}_0(Y \mid X) > \delta > 0$, F_{X0}-a.e., for some $\delta > 0$. Assume that there exists a $C_0 < \infty$ so that $\sup_Y \mid Y \mid \leq C_0$, $\sup_W \mid \psi_0(W) \mid < C_0$, and $\sup_n \sup_W \mid \psi_k(W \mid P_n) \mid \leq C_0$ with probability 1. We also assume that $\inf_X \bar{G}^0_{n,S_n}(T \mid X) \geq \delta > 0$ with probability one.*
Definitions: *Let $M_1 = \frac{8}{\delta}C_0^2$ and $M_2 = \frac{1}{\delta}16C_0^2$.*
We have

$$\max(r_1(n), r_2(n)) \leq \max(4C_0/\delta, 4C_0^2/\delta)\sqrt{E\int(\bar{G}^0_{n,S_n} - \bar{G}_0)^2(T \mid X)dF_{X0}(X)}.$$

Results: *The statements in Theorem 1 hold with the above constants M_1, M_2 and bound on $\max(r_1(n), r_2(n))$. In particular, if*

$$\frac{\max\left(E\int(\bar{G}_0 - \bar{G}^0_{n,S_n})^2(T \mid X)dF_{X0}(X), \frac{\log^2(K(n))}{np}\right)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 0,$$

then

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty.$$

Proof. Condition A1 holds at $\eta_0 = G_0$. As in example 1, it follows that condition A2 holds with $M_1^* = \frac{1}{\delta}4C_0^2$. Regarding condition A3, we note that

$$\begin{aligned}
&\int L^{*2}(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)dP_0(O) \\
&= \int \left((Y - \psi_k(\cdot \mid P^0_{n,S_n}))^2 - (Y - \psi_0(W))^2\right)^2 \frac{\Delta}{\bar{G}_0^2(T|X)}dP_0(O) \\
&= \int \left((Y - \psi_k(\cdot \mid P^0_{n,S_n}))^2 - (Y - \psi_0(W))^2\right)^2 \frac{1}{\bar{G}_0(T|X)}dF_{X0}(Y, W) \\
&\leq \frac{1}{\delta}\int \left((Y - \psi_k(\cdot \mid P^0_{n,S_n}))^2 - (Y - \psi_0(W))^2\right)^2 dF_{X0}(Y, W).
\end{aligned}$$

As in the proof of the Corollary for example 1, it follows that condition A3 holds with $M_2 = \frac{1}{\delta}16C_0^2$. This verifies conditions A1-A3.

We now define

$$D(X, \psi, \psi_0) \equiv (Y - \psi(W))^2 - (Y - \psi_0(W))^2.$$

Suppressing the dependence on G of the IPCW mapping, we have $L^*(O, \psi_k, \psi_0) = IC(O \mid D(\cdot, \psi_k, \psi_0))$. Regarding bounding $r_1(n)$, we note that for $\bar{k} \in \{\hat{k}, \tilde{k}\}$

$$\begin{aligned}
&\int (L^{*0}_{n,S_n} - L^*)(o, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(o) \\
&= \int (IC^0_{n,S_n} - IC)(O \mid D(\cdot, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)) dP_0(O) \\
&= \int D(X, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) \frac{\bar{G}_0 - \bar{G}^0_{n,S_n}}{\bar{G}^0_{n,S_n}}(T \mid X) dF_{X0}(X).
\end{aligned}$$

Thus, by the Cauchy-Schwarz inequality

$$\begin{aligned}
&E \int (IC^0_{n,S_n} - IC)(O \mid D(\cdot, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)) dP_0(O) \\
&\leq \sqrt{E \int D^2(X, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)} \sqrt{E \int \frac{(\bar{G}_0 - \bar{G}^0_{n,S_n})^2}{\bar{G}^{02}_{n,S_n}}(T \mid X) dF_{X0}(X)} \\
&\leq 4C_0 \sqrt{E \int D(X, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)} \sqrt{E \int \frac{(\bar{G}_0 - \bar{G}^0_{n,S_n})^2}{\bar{G}^{02}_{n,S_n}}(T \mid X) dF_{X0}(X)}.
\end{aligned}$$

This proves that

$$r_1(n) \leq 4C_0 \sqrt{E \int \frac{(\bar{G}_0 - \bar{G}^0_{n,S_n})^2}{\bar{G}^{02}_{n,S_n}}(T \mid X) dF_{X0}(X)}.$$

Thus $r_1(n) \leq 4C_0/\delta \sqrt{E \int (\bar{G}_0 - \bar{G}^0_{n,S_n})^2 dF_{X0}(X)}$. We also note that

$$r_2(n) = E \max_k \sqrt{\int D^2(X, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) \frac{(\bar{G}^0_{n,S_n} - \bar{G}_0)^2}{\bar{G}^{02}_{n,S_n}} \bar{G}_0(T \mid X) dF_{X0}(X)}.$$

Thus

$$r_2(n) \leq \frac{1}{\delta} \sup_X D(X, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) \sqrt{E \int (\bar{G}_0 - \bar{G}^0_{n,S_n})^2 dF_{X0}(X)},$$

where we used that for a random variable X, $E\sqrt{X} \leq \sqrt{EX}$, and that $\bar{G}^0_{n,S_n}(T \mid X) > \delta > 0$ with probability 1. Since $\sup_X D(X, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) \leq 4C_0^2$, the bound on $r_2(n)$ follows. We can now apply theorem 1 with constants M_1, M_2 and the specified bounds on $r_1(n)$ and $r_2(n)$. This completes the proof of the corollary. □

Corollary 5 *(***Survival Function Estimator Selection***)*
Setting: *Let $R(t) = I(T \leq t)$ be the indicator process for a survival time T, and $L(t)$ is a covariate process, $t \geq 0$. Let $X = \bar{X}(T) = (\bar{R}(T), \bar{L}(T)) \sim F_{X,0}$ be the full data structure of interest. Let C be a right-censoring time of T with conditional distribution $G_0(\cdot \mid X)$, given X. We assume that this conditional distribution satisfies coarsening at random: for $t \leq T$*

$$\lambda_C(t \mid X) = m(t, \bar{X}(t)) \text{ for a measurable function } m.$$

Let

$$O = (\tilde{T} = \min(T, C), \Delta = I(\tilde{T} = T), \bar{X}(\tilde{T})) \sim P_0 = P_{F_{X,0},G_0}.$$

Let $\psi_0 = E_0 B(X) \in \mathbb{R}$ for some known function $B(X)$ be the parameter of interest. For example, $\psi_0 = P(T \geq t)$ and $B(X) = I(T \geq t)$ for a given t.

Define

$$IC(O \mid G, D) = D(X)\frac{\Delta}{\bar{G}(T \mid X)}$$

as the so called inverse probability of censoring weighted mapping from full data functions $D(X)$ to observed data functions. We define

$$L(O, \psi \mid \eta_0 = G_0) = IC(O \mid G_0, L(\cdot, \psi)),$$

where $L(X, \psi) = (B(X) - \psi)^2$. Then $\psi_0 = argmin_\psi E_0 L(O, \psi \mid \eta_0)$, if $\bar{G}_0(Y \mid X) > 0$, F_{X0}-a.e. We have $d_{n(1-p)}(\hat{\psi}_k, \psi_0) = E_{S_n}(\psi_k(P^0_{n,S_n}) - \psi_0)^2$.
Assumptions: *Assume $\bar{G}_0(Y \mid X) > \delta > 0$, F_{X0}-a.e., for some $\delta > 0$. Assume that there exists a $C_0 < \infty$ so that $\sup_X \mid B(X) \mid \leq C_0$, and $\sup_n \psi_k(\cdot \mid P_n) \mid \leq C_0$ with probability 1. We also assume that $\inf_X \bar{G}^0_{n,S_n}(T \mid X) \geq \delta > 0$ with probability one.*
Definitions: *Let $M_1 = \frac{8}{\delta}C_0^2$ and $M_2 = \frac{1}{\delta}16C_0^2$. We have*

$$\max(r_1(n), r_2(n)) \leq \frac{1}{\delta}\max(4C_0, 4C_0^2)\sqrt{E\int(\bar{G}_0 - \bar{G}^0_{n,S_n})^2(T \mid X)dF_{X0}(X)}.$$

Results: *The statements in Theorem 1 hold with the above constants M_1, M_2 and bound on $\max(r_1(n), r_2(n))$. In particular, if*

$$\frac{\max\left(E\int(\bar{G}_0 - \bar{G}^0_{n,S_n})^2(T \mid X)dF_{X0}(X), \frac{\log^2(K(n))}{np}\right)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 0,$$

then

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty.$$

Proof. This is an immediate corollary of the previous corollary by defining $Y = B$ and by letting W be the empty set. □

Corollary 6 *(***Density Estimator Selection based on Right Censored Data***)*
Setting: *Let $R(t) = I(T \leq t)$ be the indicator process for a survival time T, and $L(t)$ is a covariate process, $t \geq 0$. Let $X = \bar{X}(T) = (\bar{R}(T), \bar{L}(T)) \sim F_{X,0}$ be the full data structure of interest. Let $V \subset L(0)$ be a vector of baseline covariates. Let C be a right-censoring time of T with conditional distribution $G_0(\cdot \mid X)$, given X. We assume that this conditional distribution satisfies coarsening at random: for $t \leq T$*

$$\lambda_C(t \mid X) = m(t, \bar{X}(t)) \text{ for a measurable function } m.$$

Let

$$O = (\tilde{T} = \min(T, C), \Delta = I(\tilde{T} = T), \bar{X}(\tilde{T})) \sim P_0 = P_{F_{X,0},G_0}.$$

Let $\psi_0(T, V) = f_0(T \mid V)$, where $f_0(T \mid V)$ is the conditional density of T, given V. Define

$$IC(O \mid G, D) = D(X)\frac{\Delta}{\bar{G}(T \mid X)}$$

as the so called inverse probability of censoring weighted mapping from full data functions $D(X)$ to observed data functions. We define

$$L(O, \psi \mid \eta_0 = G_0) = IC(O \mid G_0, L(\cdot, \psi)),$$

where $L(X, \psi) = -\log(\psi(T, V))$. Then $\psi_0 = argmin_\psi E_0 L(O, \psi \mid \eta_0)$, if $\bar{G}_0(Y \mid X) > 0$, F_{X0}-a.e. We have $d_{n(1-p)}(\hat{\psi}_k, \psi_0) = E_{S_n} \int \log\left(\frac{\psi_k(t,v|P^0_{n,S_n})}{\psi_0(t,v)}\right) dF_0(t, v)$.
Assumptions: *Suppose that $\bar{G}_0(Y \mid X) > \delta^* > 0$, F_{X0}-a.e., for some $\delta^* > 0$. Assume that there exists a $\delta > 0$ and $C_0 < \infty$ so that $\delta < f_0(O) \leq C_0$ and $\delta <\mid \psi_k(O \mid P_n) \mid \leq C_0$, for all n, with probability 1, for P_0-almost every O.*
Definitions: *Let $M_1 = 2\log(C_0/\delta)/\delta^*$, and let $M_2 = 4C_0/\delta\delta^*$.*

We have

$$\max(r_1(n), r_2(n)) \leq \frac{1}{\delta^*}\max(\sqrt{4C_0/\delta}, 2\log(C_0/\delta)))\sqrt{E\int(\bar{G}^0_{n,S_n} - \bar{G}_0)^2(T \mid X)dF_{X0}(X)}.$$

Results: *The statements in Theorem 1 hold with the above constants* M_1, M_2 *and bound on* $\max(r_1(n), r_2(n))$. *In particular, if*

$$\frac{\max\left(E\int(\bar{G}_0 - \bar{G}^0_{n,S_n})^2(T \mid X)dF_{X0}(X), \frac{\log^2(K(n))}{np}\right)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 0,$$

then

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty.$$

Proof. This is a copy of the proof of the previous corollary, but where we now use the M_1, M_2 from the density estimator selection example. □

Corollary 7 *(***Multivariate Predictor Selection***)*
Setting: *Let* $O = (Y = (Y_1, \ldots, Y_l), W) \sim P_0$, *where* Y *is a multivariate random outcome vector and* W *a vector of covariates. Let* $\psi_0(W) \equiv E(Y \mid W) = (E(Y_1 \mid W), \ldots, E(Y_l \mid W))$ *be the multivariate conditional expectation of* Y, *given* W. *For a candidate multivariate predictor* $\psi(W)$, *we define*

$$L(O, \psi \mid \eta_0) \equiv (Y - \psi(W))^\top \eta_0(W)(Y - \psi(W)),$$

where η_0 *is a symmetric* $l \times l$*-matrix function of* W. *If* η_0 *is a user supplied known matrix, then it is not a nuisance parameter and we can denote the loss function as* $L(O, \psi)$. *However, if* η_0 *represents a limit of an estimator of an unknown matrix such as*

$$\left[E_0\left(\{Y - E_0(Y \mid W)\}\{Y - E_0(Y \mid W)\}^\top \mid W\right)\right]^{-1},$$

then η_0 *denotes a nuisance parameter which is consistently estimated from the data. For any symmetric matrix function* $\eta_0(W)$ *we have*

$$\psi_0 = argmin_\psi E_0 L(O, \psi \mid \eta_0).$$

We have $d_{n(1-p)}(\hat{\psi}_k, \psi_0) = E_{S_n}\int \parallel \eta_0^{0.5}\left(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W)\right) \parallel^2 dF_0(W)$, *where* $\parallel \cdot \parallel$ *denotes the standard euclidean norm.*
Assumptions: *Let* $C_0 \equiv \sup_{Y,j} \mid (\eta_0^{0.5}Y)_j \mid < \infty$. *Let* $\mathbf{\Psi}(C_1) \equiv \{\psi : \sup_W \sum_{j=1}^l \mid (\eta_0^{0.5}\psi(W))_j \mid \leq C_1\}$, *and assume that* $\psi_0, \hat{\psi}_k \in \mathbf{\Psi}(C_1)$ *with probability 1. If we define* $c(W) \equiv \sup_{\parallel x \parallel = 1} \parallel \eta_0^{0.5}(W)(x) \parallel$ *as the matrix norm of the linear operator* $\eta_0^{0.5}(W) : \mathbb{R}^l \to \mathbb{R}^l$, *then, we can choose*

$C_0 = \sup_W c(W) * \sup_Y \mid Y \mid$.

Definitions: *Let* $M_1^* = 5C_0 * C_1$, *and* $M_2 = 16lC_0^2 \sup_W c(W)^2$. *If* $W =$ *(i.e., is empty), then we can set*

$$M_2 = 4 \sup_{\psi \in \mathbf{\Psi}(C_1)} \frac{E_W \parallel \Sigma^{0.5}(W)\eta^{0.5}(W)(\psi - \psi_0)(W) \parallel^2}{E_W \parallel \eta^{0.5}(W)(\psi - \psi_0)(W) \parallel^2} \leq 4 * \left\{ \sup_W \parallel \Sigma^{0.5}(W) \parallel \right\}^2,$$

where $\Sigma(W) = COV(\eta^{0.5} Y \mid W)$.
Results: *In the case* η_0 *is known, we can apply Theorem 1 with these values for* M_1 *and* M_2 *and* $r_1(n) = r_2(n) = 0$. *Thus, we have for any* $\delta > 0$

$$Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) \leq (1 + 2\delta) Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0) + 2c(M_1, M_2, \delta) \frac{1 + \log(K(n))}{np}.$$

In particular, if $\frac{\log(K(n))}{npEd_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 0$, *then*

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty.$$

Results for unknown η_0**:** *If* η_0 *is a limit of an estimator* η^0_{n,S_n}, *then we can apply Theorem 1 with these values for* M_1, M_2, *and* $r_1(n), r_2(n)$ *as specified in the Lemmas 9 and 10 in the Appendix, respectively.*

Proof. Assumptions A1-A3:
A1: Let η_0 be a limit of the estimator $\eta(P_n)$. Equation (4) tells us that $\eta_0 \in \Gamma(P_0)$ for any given limit $\eta(W)$. Thus assumption A1 holds.
A2: Firstly, we note that

$$(Y - \psi_k(W \mid P^0_{n,S_n}))^\top \eta_0(W)(Y - \psi_k(W \mid P^0_{n,S_n})) = \parallel \eta_0^{0.5}(W)(Y - \psi_k(W \mid P^0_{n,S_n})) \parallel^2,$$

where $\parallel x \parallel = \sqrt{\sum_{j=1}^l x_j^2}$ is the euclidean norm in $\mathbb{R}^l$, and $\eta_0^{0.5}$ is the square root of η_0. Thus

$$L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) = \parallel \eta_0^{0.5}(W)(Y - \psi_k(W \mid P^0_{n,S_n})) \parallel^2 - \parallel \eta_0^{0.5}(W)(Y - \psi_0(W)) \parallel^2, \tag{26}$$

where these two terms represent $L(O, \psi_k(\cdot \mid P^0_{n,S_n} \mid \eta_0)$ and $L(O, \psi_0 \mid \eta_0$, respectively. For notational convenience, we define

$$\begin{aligned} \vec{a} &= \eta_0^{0.5}(W) Y \\ \vec{b}_0 &= \eta_0^{0.5}(W)\psi_0(W) \\ \vec{b} &= \eta_0^{0.5}(W)\psi_k(W \mid P^0_{n,S_n}). \end{aligned}$$

Now, note that

$$\begin{aligned} L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) &= \sum_{j=1}^{l}(a_j - b_j)^2 - \sum_{j=1}^{l}(a_j - b_{0j})^2 \\ &= \sum_{j=1}^{l} -2a_j(b_j - b_{0j}) + b_j^2 - b_{0j}^2. \end{aligned}$$

By assumption, $\max_j a_j \leq C_0$, and $\sum_j \mid b_j \mid \leq C_1$. Thus, it follows that the last expression is bounded by $mC_0 * C_1 = M_1^*$, so that Assumption A1 holds with this value of M_1^*.

A3: Firstly, consider the case that $W =$. Above, we represented $L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)$ as $\sum -2a_j(b_j - b_{0j}) + b_j^2 - b_{0j}^2$. Notice that, given W, only a_j is random. Let $c_j \equiv 2(b_j - b_{0j})$. Then

$$\mathrm{VAR}_0(L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)) = \mathrm{VAR}(\sum_j c_j a_j) = 4E_0(b - b_0)^t \Sigma (b - b_0).$$

It is now easy to see that this is indeed bounded by by $M_2 E_0 \sum_j (b_j - b_{0j})^2$, where the M_2 is specified in the theorem.

Let's now consider the general case. Firstly, equality (26) and some straightforward algebra shows that

$$\int L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) = \int \parallel \eta_0^{0.5}(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W). \tag{27}$$

Define

$$\epsilon_k(Y, W \mid P^0_{n,S_n}) \equiv \eta_0^{0.5}(Y - \psi_k(W \mid P^0_{n,S_n}))$$

and $\epsilon_0(Y, W) \equiv \eta_0^{0.5}(Y - \psi_0(W))$. Then we have the following representation:

$$L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) = \left(\sum_{j=1}^{l} \epsilon_{k,j}^2(Y, W \mid P^0_{n,S_n}) - \epsilon_{0,j}^2(Y, W) \right).$$

We now have:

$$
\begin{aligned}
&\int L^{*2}(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&= \int \left(\sum_{j=1}^l \epsilon^2_{k,j}(O \mid P^0_{n,S_n}) - \epsilon^2_{0,j}(O) \right)^2 dP_0(O) \\
&= \int \left(\sum_{j=1}^l (\epsilon_{k,j}(O \mid P^0_{n,S_n}) - \epsilon_{0j}(O))(\epsilon_{k,j}(O \mid P^0_{n,S_n}) + \epsilon_{0j}(O)) \right)^2 dP_0(O) \\
&\leq \int \sum_{j=1}^l \left\{ \epsilon_{k,j}(O \mid P^0_{n,S_n}) - \epsilon_{0j}(O) \right\}^2 \sum_{j=1}^l \left\{ \epsilon_{k,j}(O \mid P^0_{n,S_n}) + \epsilon_{0j}(O) \right\}^2 dP_0(O) \\
&\leq \sup_O \parallel \epsilon_k(O \mid P^0_{n,S_n}) + \epsilon_0(O) \parallel^2 \int \sum_{j=1}^l \left\{ \epsilon_{k,j}(O \mid P^0_{n,S_n}) - \epsilon_{0j}(O) \right\}^2 dP_0(O) \\
&= \sup_O \parallel \epsilon_k(O \mid P^0_{n,S_n}) + \epsilon_0(O) \parallel^2 \int \parallel \eta_1^{0.5}(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W) \\
&= \sup_O \parallel \epsilon_k(O \mid P^0_{n,S_n}) + \epsilon_0(O) \parallel^2 \int L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \\
&\leq \sup_W c(W)^2 16 * l M_0^2 \int L^*(O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O),
\end{aligned}
$$

where we used at the last inequality that $\parallel a + b \parallel^2 \leq 4 \max(\parallel a \parallel^2, \parallel b \parallel^2)$. This proves that A3 holds with

$$M_2 = 16 l M_0^2 \sup_W c(W)^2.$$

We have established assumptions A1-A3. So, if η_0 is given, then we can apply Theorem 1 with $M_1 = 2M_1^*$, M_2, and $r_1(n) = r_2(n) = 0$. This proves the statements in the corollary. In the case that η_0 is estimated, the Lemmas 9 and 10 in the Appendix establish bounds for $r_1(n)$ and $r_2(n)$, respectively. This completes the proof of the corollary. □

Corollary 8 *(***Counterfactual Predictor Selection***)*
Setting: *Let $X = ((Y_a, a \in \mathcal{A}), W) \sim F_{X,0}$ be the full data structure of interest, where W denotes baseline covariates and Y_a denotes the outcome on a subject if the subject would have taken treatment a. Let A be a random variable with conditional probability distribution $g_0(a \mid X) \equiv P(A = a \mid X)$, which denotes the treatment the subject actually took. The observed data structure is*

$$O = (A, Y_A, W) \sim P_0 = P_{F_{X0}, g_0}.$$

We assume that treatment is randomized within strata of W: $g_0(a \mid X) = g_0(a \mid W)$ for all $a \in \mathcal{A}$. Let

$$\psi_0(a, V) = E(Y_a \mid V).$$

Define as full-data loss function

$$L(X, \psi) = \sum_{a \in \mathcal{A}} (Y_a - \psi(a, V))^2.$$

We have

$$\psi_0 = argmin_\psi E_{F_{X0}} L(X, \psi).$$

We will now choose as loss function the double robust mapping applied to this full data loss function (van der Laan and Robins (2002), Section 6.3):

$$\begin{aligned} L(O, \psi \mid \eta_0) &= IC(O \mid Q_0, g_0, L(\cdot, \psi)) \\ &\equiv \frac{(Y - \psi(A,V))^2}{g_0(A \mid W)} - \frac{1}{g_0(A \mid W)} E_0((Y - \psi(A,V))^2 \mid A, W) \\ &\quad + \sum_{a \in \mathcal{A}} E_0((Y - \psi(A,V))^2 \mid A = a, W). \end{aligned}$$

Here $Q_0(A,W) = (E(Y \mid A,W), E(Y^2 \mid A,W))$. *For a treatment mechanism* g_1 *satisfing the so called experimental treatment assignment assumption (ETA), that is,* $\min_{a \in \mathcal{A}} g_1(a \mid W) > 0$ P_0*-a.e., we have*

$$E_{P_0} IC(O \mid Q_1, g_1, L(\cdot, \psi)) = E_{F_{X0}} L(X, \psi) \text{ if either } g_1 = g_0 \text{ or } Q_1 = Q_0.$$

We have $d_{n(1-p)}(\hat{\psi}_k, \psi_0) = \sum_{a \in \mathcal{A}} \int (\psi_k(a, V \mid P^0_{n,S_n}) - \psi_0(a,V))^2 dF_0(V)$.
Assumptions: *Let* g_1, Q_1 *be the limits of* $g^0_{n,S_n}, Q^0_{n,S_n} = Q(F^0_{n,S_n})$, *where* F_1 *is such that* $Q_1 = Q(F_1)$ *and* $dF_1(Y \mid A,W)/dF_0(Y \mid A,W) < \infty$. *We assume that* $(g_1, Q_1) \in \Gamma(P_0)$, *where* $\Gamma(P_0)$ *consists of all* (Q, g) *with* g*'s satisfying the ETA assumption and either* $g = g_0$ *or* $Q = Q_0$.
We assume that there exists a $C_0 < \infty$ *and* $\delta > 0$ *so that* $\sup_Y \mid Y \mid < C_0$, $\sup_{a,V} \mid \psi_k(a, V \mid P_n) \mid < C_0$, $\sup_{a,V} \mid \psi_0(a,V) \mid < C_0$, $\sup_{A,W} \mid (Q_{11}, Q_{12})(A,W) \mid < (C_0, C_0^2)$, $\inf_{A,W} g_1(A \mid W) > \delta > 0$, *and* $\inf_{A,W} g^0_{n,S_n}(A \mid W) > \delta > 0$ *with probability 1.*
Definitions: *Let*

$$\begin{aligned} M_1 &= \frac{16 C_0^2}{\delta} + \mid \mathcal{A} \mid 8 C_0^2 \\ M_2 &= \frac{16 C_0^2}{\delta} \sup_O \mid \frac{dP_{F_{X0}, G_0}}{dP_{F_1, G_1}}(O) \mid . \end{aligned}$$

Define

$$\begin{aligned} h^*_n(Y,A,W) &= \frac{(g_1 - g^0_{n,S_n}) g_0}{g^0_{n,S_n} g_1}(A \mid W) + \frac{(g^0_{n,S_n} - g_1) g_0}{g^0_{n,S_n} g_1}(A \mid W) \frac{dF_1}{dF_0}(Y \mid A, W) \\ &+ \frac{g_0(A \mid W)}{g^0_{n,S_n}}(A \mid W) \frac{d(F^0_{n,S_n} - F_1)}{dF_0}(Y \mid A, W) + g_0(A \mid W) \frac{d(F^0_{n,S_n} - F_1)}{dF_0}(Y \mid A, W). \end{aligned}$$

We have

$$r_1(n) \leq 4C_0\sqrt{E\int h_n^{*2}(Y,A,W)\frac{1}{g_0(A \mid W)}dP_{F_{X0},g_0}(Y,A,W)}.$$

We also define

$$\begin{aligned} h_{1n}(A,W) \equiv & \frac{2 \mid g^0_{n,S_n} - g_1 \mid}{g^0_{n,S_n} g_1}(A \mid W) + \frac{1}{g^0_{n,S_n}(A \mid W)}\int \mid dF^0_{n,S_n} - dF_1 \mid (y \mid A,W) \\ & + \sum_{a \in \mathcal{A}} \int \mid dF^0_{n,S_n} - dF_1 \mid (y \mid A = a, W). \end{aligned}$$

We have

$$r_2(n) \leq 4C_0^2 E\sqrt{\int h_{1n}^2(A,W)dP_0(A,W)}.$$

Results: *The statements of Theorem 1 hold with these values of M_1, M_2 and bounds on $r_1(n)$ and $r_2(n)$. In particular, if*

$$\frac{\max\left(r_1(n), r_2(n), \frac{\log^2(K(n))}{np}\right)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 0,$$

then

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0)} \to 1 \text{ for } n \to \infty.$$

Proof: We need to verify assumptions A1-A3 of Theorem 1.
A1: Let g_1, Q_1 be the limits of g^0_{n,S_n}, Q^0_{n,S_n}. We assumed that $(g_1, Q_1) \in \Gamma(P_0)$, where $\Gamma(P_0)$ consists of all (Q, g) with g's satisfying the ETA assumption and either $g = g_0$ or $Q = Q_0$.
A2: Given our assumptions, straightforward algebra shows now that A2 holds with

$$M_1^* = \frac{8C_0^2}{\delta} + \mid \mathcal{A} \mid 4C_0^2.$$

A3: Define $D(X, \psi, \psi_0) = L(X, \psi) - L(X, \psi_0)$. We also define $IC_0(O \mid G_0, D(\cdot, \psi, \psi_0) = \{(Y - \psi(A,V))^2 - (Y - \psi_0(A,V))^2\}/g_0(A \mid W)$. In general, we have the following result.

Lemma 7 *For any (F_1, G_1) for which $P_0 \ll P_{F_1,G_1}$ we have the following inequality*

$$\begin{aligned}
&\int IC^2(O \mid Q(F_1), G_1, D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0))dP_0(O) \\
&= \int IC^2(O \mid Q(F_1), G_1, D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0))\frac{dP_{F_{X0},G_0}}{dP_{F_1,G_1}}(O)dP_{F_1,G_1}(O) \\
&\leq \sup_O \mid \frac{dP_{F_{X0},G_0}}{dP_{F_1,G_1}}(O) \mid \int IC^2(O \mid Q(F_1), G_1, D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0))dP_{F_1,G_1}(O) \\
&\leq \sup_O \mid \frac{dP_{F_{X0},G_0}}{dP_{F_1,G_1}}(O) \mid \int IC_0^2(O \mid G_1, D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0))dP_{F_1,G_1}(O).
\end{aligned}$$

In the same manner as in the previous example, it follows that if $(Q(F_1), G_1) \in \Gamma(P_0)$, then

$$\begin{aligned}
&\int IC_0^2(O \mid G_1, D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0))dP_{F_1,G_1}(O) \\
&\leq \tfrac{16C_0^2}{\delta} \int IC_0(O \mid G_1, D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0))dP_{F_1,G_1}(O) \\
&= \tfrac{16C_0^2}{\delta} \int IC(O \mid G_1, D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0))dP_{F_1,G_1}(O).
\end{aligned}$$

Combining the latter bound with Lemma 7 teaches us that A3 holds with

$$M_2 = \frac{16C_0^2}{\delta} \sup_O \mid \frac{dP_{F_{X0},G_0}}{dP_{F_1,G_1}}(O) \mid .$$

We have now verified assumptions A1-A3. It remains to establish the claimed bounds on $r_1(n)$ and $r_2(n)$. Define $D_k(Y, A, V) \equiv (Y - \psi_k(A, V \mid P^0_{n,S_n}))^2 - (Y - \psi_0(A, V))^2$. Regarding bounding $r_1(n)$, we have

$$\begin{aligned}
&\int (IC^0_{n,S_n} - IC)(O \mid D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)dP_0(O) \\
&= E_W \textstyle\sum_a \int D_k(Y_a, a, V)dF_0(Y_a \mid W)\frac{g_1 - g^0_{n,S_n})}{g^0_{n,S_n} g_1}(a \mid W)g_0(a \mid W) \\
&+ E_W \textstyle\sum_a \int D_k(Y_a, a, V)dF_1(Y_a \mid W)\frac{g^0_{n,S_n} - g_1}{g^0_{n,S_n} g_1}(a \mid W)g_0(a \mid W) \\
&+ E_W \textstyle\sum_a \int D_k(Y_a, a, V)d(F^0_{n,S_n} - F_1)(Y_a \mid W)\frac{g_0}{g^0_{n,S_n}}(a \mid W) \\
&+ E_W \textstyle\sum_a \int D_k(Y_a, a, V)d(F^0_{n,S_n} - F_1)(Y_a \mid W)g_0(a \mid W) \\
&\equiv E_{F_{X0},g_0} D_k(Y_A, A, V)h_n(X, A),
\end{aligned}$$

where

$$\begin{aligned}
h_n(X, A) \;=\; &\frac{g_1 - g^0_{n,S_n}}{g^0_{n,S_n} g_1}(A \mid W) + \frac{g^0_{n,S_n} - g_1}{g^0_{n,S_n} g_1}(A \mid W)\frac{dF_1}{dF_0}(Y \mid A, W) \\
&+ \frac{1}{g^0_{n,S_n}}(A \mid W)\frac{d(F^0_{n,S_n} - F_1)}{dF_0}(Y \mid A, W) + \frac{d(F^0_{n,S_n} - F_1)}{dF_0}(Y \mid A, W).
\end{aligned}$$

We now write

$$E_{F_{X0},g0}D_k(Y_A,A,V)h_n(X,A) = E_{F_{X0},g^*}D_k(Y_A,A,V)h_n^*(X,A),$$

where $g^*(\cdot \mid W)$ is the counting measure on $\mathcal{A}$ and

$$h_n^*(X,A) \equiv h_n(X,A)g_0(A \mid W).$$

Application of the Cauchy-Schwarz inequality yields now

$$EE_{F_{X0},g^*}D_k(Y_A,A,V)h_n^*(X,A) \leq \sqrt{EE_{F_{X0},g^*}D_k^2(Y_A,A,V)}\sqrt{EE_{F_{X0},g^*}h_n^{*2}(X,A)}.$$

We now note that

$$\begin{aligned} EE_{F_{X0},g^*}D_k^2(Y_A,A,V) &= EE_{F_{X0}}\left(\sum_{a\in\mathcal{A}} D_k(Y_a,a,V)\right)^2 \\ &\leq EE_{F_{X0}} \mid \mathcal{A} \mid \sum_{a\in\mathcal{A}} D_k^2(Y_a,a,V) \\ &\leq 16C_0^2 \sum_{a\in\mathcal{A}} EE_{F_{X0}}D_k(Y_a,a,V). \end{aligned}$$

This proves the following lemma.

Lemma 8 *Define*

$$\begin{aligned} h_n^*(Y,A,W) &= \frac{g_0(g_1-g_{n,S_n}^0)}{g_{n,S_n}^0 g_1}(A \mid W) + \frac{g_0(g_{n,S_n}^0-g_1)}{g_{n,S_n}^0 g_1}(A \mid W)\frac{dF_1}{dF_0}(Y \mid A,W) \\ &+\frac{g_0(A \mid W)}{g_{n,S_n}^0}(A \mid W)\frac{d(F_{n,S_n}^0-F_1)}{dF_0}(Y \mid A,W) + g_0(A \mid W)\frac{d(F_{n,S_n}^0-F_1)}{dF_0}(Y \mid A,W). \end{aligned}$$

We have

$$r_1(n) \leq 4C_0\sqrt{E\int h_n^{*2}(Y,A,W)\frac{1}{g_0(A \mid W)}dP_{F_{X0},g_0}(Y,A,W)}.$$

The bound on $r_2(n)$ is derived similarly. This completes the proof of the corollary. □

Lemma for Corollary 7.

In order to bound $r_1(n)$ we prove the following lemma.

Lemma 9 *We recall (27). We have*

$$E \int L^{*0}_{n,S_n}(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) - L^*(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \leq$$

$$\sqrt{E \int \parallel (\eta^{0,0.5}_{n,S_n} - \eta^{0.5}_0)(W)(\psi_{\bar{k}}(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W)}$$

$$\times \sqrt{E \int \parallel (\eta^{0,0.5}_{n,S_n} + \eta^{0.5}_0)(W)(\psi_{\bar{k}}(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W)}.$$

Thus, if for a constant ϵ_n

$$\parallel \eta^{0,0.5}_{n,S_n}(W)(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel < (1+\epsilon_n) \parallel \eta^{0.5}_0(W)(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel,$$

with probability 1, then

$$E \int L^{*0}_{n,S_n}(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) - L^*(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) \leq$$

$$\sqrt{E \int \parallel (\eta^{0,0.5}_{n,S_n} - \eta^{0.5}_0)(W)(\psi_{\bar{k}}(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W)}$$

$$\times \sqrt{(4 + 4\epsilon_n + \epsilon_n^2) E \int \parallel \eta^{0.5}_0(W)(\psi_{\bar{k}}(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W)}.$$

Thus, in this case

$$r_1(n) \leq \max_{\bar{k} \in \{\hat{k}, \tilde{k}\}} \sqrt{E \int \parallel (\eta^{0,0.5}_{n,S_n} - \eta^{0.5}_0)(W)(\psi_{\bar{k}}(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W)}$$
$$\times \sqrt{4 + 4\epsilon_n + \epsilon_n^2}.$$

Proof. Let a_j, $a^{0.5}_{j,n,S_n}$, be the j-th row-vector of $\eta^{0.5}_0$ and $\eta^{0.5}_{n,S_n}$, respectively, $j = 1, \ldots, l$. Then, using short-hand notation

$$E \int L^{*0}_{n,S_n}(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) - L^*(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O) =$$

$$E \int \textstyle\sum_j \{a^{0\top}_{j,n,S_n}(\psi_{\bar{k}} - \psi_0)\}^2 - \{a^\top_j(\psi_{\bar{k}} - \psi_0)\}^2 dF_0(W)$$

$$= E \int \textstyle\sum_j (a^{0\top}_{j,n,S_n} - a^\top_j)(\psi_{\bar{k}} - \psi_0)(a^{0\top}_{j,n,S_n} + a^\top_j)(\psi_{\bar{k}} - \psi_0) dF_0(W).$$

Subsequently, we apply Cauchy-Schwarz inequality in the following manner:

$$E_{P_n} \int_W \sum_j f(j,W,P_n)g(j,W,P_n)dF_0(W) \leq \sqrt{E_{P_n} \int_W \sum_j f^2(j,W,P_n)dF_0(W)}$$
$$\times \sqrt{E_{P_n} \int_W \sum_j g^2(j,W,P_n)dF_0(W)}.$$

This yields precisely the claimed inequality. The other statements in the lemma are trivially verified. □

In the same manner we establish a bound for $r_2(n)$, as given in the following lemma.

Lemma 10 *We have*

$$\sqrt{\int \left(L^{*0}_{n,S_n}(O,\psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}),\psi_0) - L^*(O,\psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}),\psi_0)\right)^2 dP_0(O)}$$

$$\leq \sup_W \parallel (\eta^{0.5}_{n,S_n} + \eta^{0.5}_0)(W)(\psi_{\bar{k}}(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel \times$$

$$\sqrt{\int \parallel (\eta^{0.5}_{n,S_n} + \eta^{0.5}_0)(W)(\psi_{\bar{k}}(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W)}.$$

Thus, if

$$\max_k \sup_W \parallel (\eta^{0.5}_{n,S_n} + \eta^{0.5}_0)(W)(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel \leq M < \infty \text{ with probability 1,}$$

then

$$r_2(n) \leq ME \max_k \sqrt{\int \parallel (\eta^{0.5}_{n,S_n} + \eta^{0.5}_0)(W)(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W)) \parallel^2 dF_0(W)}.$$

Selection With Censored Data

8 The cross-validation selector

Our censored data examples are special cases of the following complete general strategy for generalizing selection procedures one would use based on uncensored data to selection procedures based on censored data. Note that the cross-validation selection procedure $\hat{k}$ is defined once we find the appropriate loss function $L(O, \psi \mid \eta_0)$.

Let $X \sim F_{X,0}$ be a full data structure of interest. Let $\psi_0(\cdot) = \psi(\cdot \mid F_{X,0})$ be a parameter (function) of $F_{X,0}$ of interest. Let the parameter space be $\mathbf{\Psi}$. Let $(X, \psi) \rightarrow L(X, \psi) \in \mathbb{R}$ be a "loss function" whose expectation is minimized by ψ_0:

$$\begin{aligned} \psi_0 &= \text{argmin}_{\psi \in \mathbf{\Psi}} \int L(X, \psi) dF_{X,0}(X) \\ &= \text{argmin}_{\psi \in \mathbf{\Psi}} E_0 L(X, \psi). \end{aligned}$$

In real life applications, we often do not observe the full data X but a censored version. Let C denote a censoring variable. We will represent the *observed data random variable* with $O = \Phi(C, X)$ for some known function Φ. The distribution $P_0 = P_{F_{X0}, G_0}$ of the observed data O is indexed by the full data distribution F_{X0} and the conditional probability distribution $G_0(\cdot \mid X)$ of the censoring variable C given X. We refer to $G_0(\cdot \mid X)$ as the censoring mechanism and sometimes simply denote it with G_0.

We assume *coarsening at random* (CAR) on the censoring mechanism. Coarsening at random was originally formulated by Heitjan and Rubin (1991) and further generalized by Jacobsen and Keiding (1995) and Gill et al. (1997). We refer to Robins and Rotnitzky (1992) and Robins (1993) for the introduction and discussion of this CAR definition for the right censored data structure, and, in general, we refer to van der Laan and Robins (2002) for the definitions of CAR for various censored data structures. Let $\mathcal{G}(CAR)$ be the set of all conditional distributions $G(\cdot \mid X)$ satisfying CAR. Because of CAR, we have that the density of the observed data distribution $P_{F_X,G}$ w.r.t. an appropriate dominating measure (Gill et al. (1997)) factorizes into an F_X and G factor:

$$\frac{dP_{F_X,G}}{d\mu}(O) = Q_{F_X}(O) g_{O|X}(O \mid X),$$

where $g(O \mid X)$ is a density of the conditonal distribution of O, given X, w.r.t. to a dominating measure satisfying CAR itself (Gill et al. (1997)). Note that $g_{O|X}$ depends on G only.
Given an empirical distribution P_n based on an i.i.d. sample $\{O_i, i = 1, \cdots, n\}$ of size n, let $\psi_k(. \mid P_n), k \in \{1, \cdots, K(n)\}$ be well defined estimators of the parameter ψ_0. We will not discuss the available methods to obtain such estimators for different full data distribution models and censoring mechanism models in this paper. We refer the reader to van der Laan and Robins (2002) for a comprehensive and in depth presentation of inverse probability of censoring weighted estimators and doubly robust locally efficient estimators of the full data parameters in (multivariate) generalized linear regression and multiplicative intensity models for the full data distribution, for a class of censored data structures, including right-censored data, multivariate right-censored data, cross-sectional data, and missing data structures occurring in causal inference. As in our general presentation of the methodology in Section 1, we are concerned with selecting a $\hat{k}$ among $\{1, \cdots, K(n)\}$ such that

$$d_n(\hat{\psi}_{\hat{k}}, \psi_0) = \int L(X, \psi_{\hat{k}}(\cdot \mid P_n)) - L(X, \psi_0)) dF_{X0}(X)$$

converges to zero asymptotically as fast as

$$d_n(\hat{\psi}_{\tilde{k}}, \psi_0) = \min_{k \in \{1,\ldots,K(n)\}} d_n(\hat{\psi}_k, \psi_0).$$

Note that we cannot directly apply the selector $\hat{k}$ of Section 1 since the loss funtion $L(X, \psi)$ is not a function of the observed data O.

Let $\mathcal{D}$ be a set of full data functions so that $P(L(\cdot, \psi_k(\cdot \mid P_n)) \in \mathcal{D}) = 1$ for all $k = 1, \ldots,$. The fundamental idea is to replace the full data function $L(X, \psi)$ by a function $L(O, \psi \mid \eta_0)$ of the observed data which has the same expectation, and apply the general method $\hat{k}$ of Section 1. It appears that the censored data methodology on estimating functions as presented in van der Laan and Robins (2002) yields for each censored data structure a mapping from a full data function $L(X, \psi)$ to an observed data function which has the same expectation. Consequently, a simple referal to this work will provide us with the wished selection methods based on censored data. Another way to think about this is that we consider the risk of a given estimator $\hat{\psi}$ based on the training sample, that is, the expectation under F_{X0} of the loss $L(X, \psi(\cdot \mid P^0_{n,S_n}))$, as a full data parameter of interest and apply methods as presented in van der Laan and Robins (2002) for estimating it consistently and efficiently using the observed data.

The general estimating function methodology as presented in van der Laan and Robins (2002) maps a given full data function $D \in \mathcal{D}$ into an observed data function $IC[O \mid Q_0, G_0, D]$ indexed by nuisance parameters $Q_0 = Q(F_{X0}, G_0)$ and G_0. We define $\Gamma(P_{F_{X0},G_0})$ as the set of all pairs (F_1, G_1), $G_1 \in \mathcal{G}(CAR)$, for which for all $D \in \mathcal{D}$

$$E_{P_0} IC[O \mid Q(F_1, G_1), G_1, D] = E_{F_{X0}} D(X).$$

In other words, $IC[O \mid Q(F_{X0}, G_0), G_0, D]$ is a function of the observed data which has the same expectation as the unobserved full data function $D(X)$, as long as $(F_1, G_1) \in \Gamma(P_0)$. We can now apply our general cross-validation selection method $\hat{k}$ as defined in Section 1 with

$$L(O, \psi \mid \eta_0 = (F_1, G_1)) = IC[O \mid Q(F_1, G_1), G_1, L(\cdot, \psi)], \tag{28}$$

for any $(F_1, G_1) \in \Gamma(P_0)$. Note that, by definition of $\Gamma(P_0$, we indeed have for all $\eta_0 \in \Gamma(P_0)$

$$\psi_0 = \text{argmin}_\psi E_{P_0} L(O, \psi \mid \eta_0).$$

The estimating function methodology as presented in van der Laan and Robins (2002) provides us with essentially two mappings $D \rightarrow IC[O \mid Q, G, D]$ from full data estimating functions to observed data estimating functions: Inverse Probability of Censoring Weighted (IPCW) Estimating Functions and Double Robust IPCW estimating functions. Both classes are the range of a (typically) linear mapping $D \rightarrow IC[O \mid Q_0, G_0, D]$ applied to full-data estimating functions $D(X)$. The Inverse Probability of Censoring Weighted estimating functions are only indexed by G and are thus such that $\Gamma(P_0) = \{(F, G_0) : F\}$, where F ranges over all possible full data distributions. Here G_0 has to also satisfy a particular support condition necessary for making the estimating function $IC[O \mid Q_0, G_0, D]$ unbiased for $E_0 D(X)$ under P_0. This support condition can be considered as an identifiability condition: see van der Laan and Robins (2002). The double robust estimating functions as defined and developed in van der Laan and Robins (2002) are such that $\Gamma(P_0) = \{(F_1, G_1) : F_1 = F_{X0} \text{ or } G_1 = G_0\}$, where, again, (the possibly misspecified) $G_1 \in \mathcal{G}(CAR)$ needs to satisfy the same support (identifiability) condition. In words, the IPCW estimating functions require correct specification of the censoring mechanism G_0, while the double robust estimating functions require either correct specification of G_0 or correct specification of F_{X0}. We refer to censored data examples 3 and 4 for a particular

IPCW observed data function and to example 6 for a double robust IPCW estimating function, and, note that in both examples we specify the required support condition on the possibly misspecified G_1.

Let $S_n \in \{0,1\}^n$ be a random vector independent of P_n. A realization of S_n defines a particular split of the sample of n observations into a training sample $\{i \in \{1,\cdots,n\} : S_{n,i} = 0\}$ and a validation sample $\{i \in \{1,\cdots,n\} : S_{n,i} = 1\}$. Let P^0_{n,S_n}, P^1_{n,S_n} denote the empirical distributions of the training and the validation sample, respectively. Let the proportion $p(n) \equiv p = 1/n \sum_{i=1}^n S_{n,i} \in (0,1)$ of observations in the validation sample be constant (i.e, non-random).

Let $G^0_{n,S_n}(. \mid X)$ and Q^0_{n,S_n} be estimators of $G_0(. \mid X)$ and $Q(F_{X0}, G_0)$, respectively, based on the training sample. We define the following cross-validated risk estimate of the true conditional risk $\tilde{\theta}_{n(1-p)}(k) \equiv E_{S_n} \int L(X, \psi_k(\cdot \mid P^0_{n,S_n}))dF_{X0}$:

$$\begin{aligned}
\hat{\theta}_{n(1-p)}(k) &= E_{S_n} \int L(o, \psi_k(\cdot \mid P^0_{n,S_n} \mid \eta_0)dP_0(o) \\
&= E_{S_n} \int IC[O \mid Q^0_{n,S_n}, G^0_{n,S_n}, L(., \psi_k(. \mid P^0_{n,S_n}))]dP^1_{n,S_n}(O) \\
&= E_{S_n} \frac{1}{np} \sum_{i=1}^n I(S_n(i) = 1) IC[O_i \mid Q^0_{n,S_n}, G^0_{n,S_n}, L(., \psi_k(\cdot \mid P^0_{n,S_n}))].
\end{aligned}$$

This risk estimate defines the cross-validation selector $\hat{k}$ of k given by

$$\begin{aligned}
\hat{k} &= \text{argmin}_{k \in \{1,\cdots,K(n)\}} \hat{\theta}_{n(1-p)}(k) \\
&= \text{argmin}_{k \in \{1,\ldots,K(n)\}} E_{S_n} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) dP^1_{n,S_n}(O).
\end{aligned}$$

Note, this is equivalent with the general formula for our selector $\hat{k}$ corresponding with the choice of loss function (28):

$$\hat{k} = \text{argmin}_{k \in \{1,\cdots,K(n)\}} E_{S_n} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) dP^1_{n,S_n}(O).$$

9 Theorems for Selection with Censored Data

We will benchmark the selector $\hat{k}$ by the minimizer of the true conditional risk function

$$\tilde{\theta}_{n(1-p)}(k) = E_{S_n} \int L(X, \psi_k(\cdot \mid P^0_{n,S_n}))dF_{X0}(X). \tag{29}$$

This quantaty equals the true conditional risk of the estimator based on $n(1-p)$ observations. The minimizer

$$\tilde{k} \;=\; min^{-1}_{k\in\{1,\cdots,K(n)\}}\tilde{\theta}_{n(1-p)}(k)$$

of the true conditional risk function for a given P_n defines a best possible choice for $\hat{k}$ since, for each given data set P_n, it indexes the best estimator among $\psi_k(.\mid P_{n(1-p)}), k \in \{1,\cdots,K(n)\}$ that achieves the optimal conditional risk based on $n(1-p)$ observations. Again, we note that this $\tilde{k}$ is different from the minimizer $\tilde{k}_n$ of $\tilde{\theta}_n(k)$ as defined in Section 1. In practice, we do not know the true conditional risk function $\tilde{\theta}_{n(1-p)}(\cdot)$ since it depends on the true distribution F_{X0}. Consequently, we do not have $\tilde{k}$ available to us.
For any $\eta_0 = (F_1, G_1) \in \Gamma(P_{F_{X0},G_0})$, we have

$$\begin{aligned}\tilde{\theta}_{n(1-p)}(k) &= E_{S_n}\int IC[O \mid Q(F_1,G_1), G_1, L(.,\psi_k(.\mid P^0_{n,S_n}))]dP_0(O)\\ &= E_{S_n}\int L(O,\psi_k(\cdot\mid P^0_{n,S_n})\mid\eta_0)dP_0(O).\end{aligned}$$

Similarly, the optimal risk

$$\theta_{opt} \equiv E_0 L(X,\psi_0)]$$

can be represented as

$$\begin{aligned}\theta_{opt} &= \int IC[O\mid Q(F_1,G_1),G_1,L(.,\psi_0)]dP_0(O)\\ &= \int L(O,\psi_0\mid\eta_0)dP_0(O)\\ &\quad \text{for any } \eta_0=(F_1,G_1)\in\Gamma(P_{F_{X0},G_0}).\end{aligned}$$

In the next subsections we state the analogues of our general Theorems 1 and 2. These results are almost direct corollaries of our general theorems, except that we provide some general strategies for verifying the Assumption A3 and for bounding $r_1(n)$. Theorem 3 applies to full data loss functions whose optimal risk θ_{opt} can be estimated at a quadratic rate, while Theorem 4 applies to general loss functions.

9.1 Quadratic loss function.

Notation. Throughout the following theorem we introduce and use the following notation:

$$D(X,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0) = L(X,\psi_k(\cdot \mid P^0_{n,S_n})) - L(X,\psi_0)$$
$$IC[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)] = IC[O \mid Q(F_1,G_1),G_1,D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)]$$
$$IC^0_{n,S_n}[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)] = IC[O \mid Q^0_{n,S_n},G^0_{n,S_n},D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)]$$
$$(IC^0_{n,S_n} - IC)[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)]$$
$$= IC^0_{n,S_n}[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)] - IC[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)].$$

Because $(IC^0_{n,S_n} - IC)$ is typically a linear real valued mapping in $D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0) \in (L^2(F_X),\langle\cdot,\cdot\rangle_{F_X})$, the Riesz-Representation theorem teaches us that the handy (i.e, unnecessary, but it yields a simple way of bounding $r_1(n)$) condition E1 in the next theorem often holds.

Theorem 3
Assumptions.
A1. $(F_1,G_1) \in \Gamma(P_0)$.
A2. *There exist a* $M_1^* < \infty$ *so that for all* k

$$\sup_O IC[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)] \leq M_1^* \ a.s.,$$

where the supremum is taken over a support of the distribution P_0 *of* O.
A3. *There exist a* $M_2 < \infty$ *so that for all* k

$$\int IC^2[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)]dP_0(O) \leq M_2 \int IC[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)]dP_0(O) \tag{30}$$

*It will typically be convenient to replace it by the following sufficient conditions A*3.1 *and A*3.2*:*
A3.1. *There exist* $M_2^* < \infty$ *so that for all* k

$$\int IC^2[O \mid D(\cdot,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)]dP_0(O) \ \leq \ M_2^* \int D^2(X,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)dF_{X,0}(X)$$

A3.2. *There exists a* M *so that*

$$\int D^2(X,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)dF_{X,0}(X) \leq M \int D(X,\psi_k(\cdot \mid P^0_{n,S_n}),\psi_0)dF_{X,0}(X).$$

Then (30) holds with $M_2 = MM_2^*$.

Definitions. *We define the following constants:*

$$\begin{aligned} M_1 &= 2M_1^* \\ c(M_1, M_2, \delta) &= 2(1+\delta)^2 \left(\frac{M_1}{3} + \frac{M_2}{\delta} \right) \\ a_0 &\equiv 2M_1/3 \\ M_3 = M_3(n) &= 3a_0 + \sqrt{2}\frac{\sqrt{\log(2)}}{\log(K(n))} + \frac{\sqrt{2}}{\sqrt{\log(K(n))}} + b_0 + \int_{b_0}^{\infty} 2K(n)^{1-m(x)}dx, \end{aligned}$$

where b_0 *is the smallest constant larger than the solution of* $1 - m(x) = 0$ *with* $m(x) = 0.5\frac{x^2}{1/\log(K(n))+a_0 x}$. *We note that* $M_3(n) \downarrow$ *in* n. *We also define the following sequences in* n*:*

$$\begin{aligned} r_1(n) &\equiv \max_{\bar{k}\in\{\hat{k},\tilde{k}\}} \frac{E\int (IC^0_{n,S_n} - IC)[O \mid D(\cdot, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)]dP_0(O)}{\sqrt{E\int IC[O \mid D(\cdot, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)]dP_0(O)}} \\ r_2(n) &\equiv E\max_k \sqrt{\int (IC^0_{n,S_n} - IC)^2[O \mid D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)]dP_0(O)} \\ \tilde{r}(n) &\equiv \sqrt{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}}. \end{aligned}$$

A method for bounding $r_1(n)$**.** *Consider the following assumption:*

E1 *For all* k *(recall* $O = \Phi(C, X)$*)*

$$\begin{aligned} &\int (IC^0_{n,S_n} - IC)[\Phi(C,X) \mid D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)]dG_0(C \mid X)dF_{X0}(X) \\ &= \int D_k(x \mid P^0_{n,S_n}) f_n(x \mid S_n, P^0_{n,S_n})dF_{X0}(x) \end{aligned}$$

for some function $f_n(\cdot \mid S_n, P^0_{n,S_n}) \in L^2(F_{X0})$. *If* **E1** *and A3.2 hold, then we have*

$$r_1(n) \leq \sqrt{M}\sqrt{E_{S_n, P^0_{n,S_n}} \int f_n^2(x \mid S_n, P^0_{n,S_n})dF_{X0}(x)}. \tag{31}$$

Finally, for any $\delta > 0$ *we define*

$$\begin{aligned} \epsilon_n(\delta) &\equiv (1+2\delta)\tilde{r}^2(n) + 2c(M_1, M_2, \delta)\frac{1+\log(K(n))}{np} \\ &+(1+\delta)r_1(n)\tilde{r}(n) + \frac{2M_3(1+\delta)\log(K(n))}{(np)^{0.5}} \max(r_2(n), (np)^{-0.5} I(r_2(n) > 0)). \end{aligned}$$

Finite Sample Result. *For any $\delta > 0$, we have*

$$\sqrt{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}} \leq \frac{r_1(n)(1+\delta) + \sqrt{r_1(n)^2(1+\delta)^2 + 4\epsilon_n(\delta)}}{2}. \quad (32)$$

Asymptotic Implication. *For any $\delta > 0$*

$$E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq (1+2\delta)\left\{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\right\} + O(H(n)),$$

where

$$H(n) \equiv \max\left(\frac{\log(K(n))}{np}, \frac{\log(K(n))r_2(n)}{(np)^{0.5}}, r^2(n), r_1(n)\tilde{r}(n), r_1(n)^{1.5}\tilde{r}(n)^{0.5}, \frac{\sqrt{\log(K(n))}r_1(n)}{(np)^{0.5}}, \frac{\sqrt{\log(K(n)}r_2(n)^{0.5}r_1(n)}{(np)^{0.25}}\right).$$

If $\frac{\max(r_1(n),r_2(n))}{\tilde{r}(n)} \to 0$ for $n \to \infty$, then

$$H(n) = O\left(\frac{\log(K(n))}{np}\right) + o\left(\max\left(\frac{\log(K(n))\tilde{r}(n)}{(np)^{0.5}}, \tilde{r}^2(n), \frac{\sqrt{\log(K(n))}}{(np)^{0.25}}\tilde{r}(n)^{1.5}\right)\right).$$

Thus, if also $\frac{\log(K(n))}{(np)\tilde{r}(n)^2} \to 0$ for $n \to \infty$, then

$$H(n) = o(\tilde{r}(n)^2) = o(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}).$$

Asymptotic Optimality. *Thus, if $\frac{\max(r_1(n),r_2(n))}{\tilde{r}(n)} \to 0$ and $\frac{\log(K(n))}{(np)\tilde{r}(n)^2} \to 0$, then*

$$\frac{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 1 \text{ for } n \to \infty. \quad (33)$$

Finally, if $\frac{\max(r_1(n),r_2(n))}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 0$ in probability, and $\frac{\log(K(n)}{(np)\{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\}} \to 0$ in probability, then

$$\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 1 \text{ in probability for } n \to \infty. \quad (34)$$

This theorem is a direct corollary of Theorem 1 and the following lemma which establishes the bound (31).

Lemma 11 *Let $\bar{k} \in \{\hat{k}, \tilde{k}\}$ Assume there exists a M so that for*

$$\int D^2(X, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dF_{X,0}(X) \leq M \int D(X, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dF_{X,0}(X).$$

In addition, assume that

$$\begin{aligned} &\int (IC^0_{n,S_n} - IC)[O(c,x) \mid D(\cdot, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)] dG_0(c \mid x) dF_{X0}(x) \\ &= \int D(X, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) f_n(x \mid S_n, P^0_{n,S_n}) dF_{X0}(x) \end{aligned}$$

for some function $f_n(\cdot \mid S_n, P^0_{n,S_n}) \in L^2(F_{X0})$. Let

$$q(n) \equiv E \int (IC^0_{n,S_n} - IC)[O \mid D(\cdot, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)] dP_0(O).$$

Then

$$q(n) \leq \sqrt{M} \sqrt{E \int f_n^2(x \mid S_n, P^0_{n,S_n}) dF_{X0}(x)} \sqrt{E \tilde{\theta}_{n(1-p)}(\bar{k}) - \theta_{opt})}.$$

Proof. Let $\bar{k} = \hat{k}$. Let B_n denote the random variable (S_n, P_n) and represent $D(x, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)$ as $D(x \mid B_n)$. Similarly, we represent $f_n(x \mid S_n, P^0_{n,S_n}) = f_n(x \mid B_n)$. The expectation of $\int D(x \mid B_n) f_n(x \mid B_n) dF_{X0}(x)$ w.r.t. B_n can be written as $\int D(x \mid b) f_n(x,b) dF_{X0}(x) dQ_n(b)$, where Q_n is the probability distribution of B_n. One can now apply the Cauchy-Schwarz inequality:

$$\begin{aligned} &\int D(x \mid b) f_n(x,b) dF_{X0}(x) dQ_n(b) \leq \\ &\sqrt{\int D^2(x \mid b) dF_{X0}(x) dQ_n(b)} \sqrt{\int f_n^2(x,b) dF_{X0}(x) dQ_n(b)}. \end{aligned}$$

By assumption,

$$\begin{aligned} \int D^2(x \mid b) dF_{X0}(x) &= \int \left\{ L(X, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n})) - L(X, \psi_0) \right\}^2 dF_{X0}(X) \\ &\leq M \int \left\{ L(X, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n})) - L(X, \psi_0) \right\} dF_{X0}(X). \end{aligned}$$

Thus

$$\sqrt{\int D^2(x \mid b) dF_X(x) dQ_n(b)} \leq \sqrt{M} \sqrt{E \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}.$$

Thus this shows that

$$E\int(IC^0_{n,S_n} - IC)[O \mid D(\cdot, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)]dP_0(O) \leq$$
$$\sqrt{M}\left(\sqrt{E\int f_n^2(x \mid S_n, P^0_{n,S_n})dF_{X0}(x)}\sqrt{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})}\right).\Box$$

9.2 General loss functions.

Theorem 4
Assumptions.
A1. $(F_1, G_1) \in \Gamma(P_0)$.
A2. *There exist a* $M_1^* < \infty$ *so that for all* k

$$\sup_O IC[O \mid D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)] \leq M_1^* \ a.s.,$$

where the supremum is taken over a support of the distribution P_0 *of* O.

Definitions. *We define the following sequences in* n*:*

$$\begin{aligned} f(M_1^*, K(n), np) &\equiv 4M_1^{*2}\sqrt{\frac{\log K(n)}{np}} \\ r_1(n) &\equiv \max_{\bar{k}\in\{\hat{k},\tilde{k}\}} \frac{E\int(IC^0_{n,S_n} - IC)[O \mid D(\cdot, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)]dP_0(O)}{\sqrt{E\int IC[O \mid D(\cdot, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0)]dP_0(O)}} \\ \tilde{r}(n) &\equiv \sqrt{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}}. \end{aligned}$$

A method for bounding $r_1(n)$**.** *Consider the following two conditions* $E1$ *and* $E2$*:*
E1. *For all* k *(recall* $O = \Phi(C, X)$*)*

$$\int(IC^0_{n,S_n} - IC)[\Phi(C, X) \mid D(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)]dG_0(C \mid X)dF_{X0}(X)$$
$$= \int D_k(x \mid P^0_{n,S_n})f_n(x \mid S_n, P^0_{n,S_n})dF_{X0}(x)$$

for some function $f_n(\cdot \mid S_n, P^0_{n,S_n}) \in L^2(F_{X0})$,
E2. *There exists a* M *so that for all* k

$$\int D^2(\cdot, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)dF_{X,0}(X) \leq M\int D(X, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0)dF_{X,0}(X).$$

If E1 and E2 hold, then we have

$$r_1(n) \leq \sqrt{M}\sqrt{E_{S_n,P^0_{n,S_n}} \int f_n^2(x \mid S_n, P^0_{n,S_n}) dF_{X0}(x)}.$$

Finally, we define

$$\epsilon_n \equiv \tilde{r}^2(n) + f(M_1^*, K(n), np) + r_1(n)\tilde{r}(n).$$

Finite Sample Result. *We have*

$$\sqrt{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}} \leq \frac{r_1(n) + \sqrt{r_1(n)^2 + 4\epsilon_n}}{2}. \tag{35}$$

Asymptotic Implication. *We have*

$$E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \;\leq\; E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} + O(H(n)),$$

where

$$H(n) \;\equiv\; \max\left(\frac{\log^{0.5}(K(n))}{(np)^{0.5}}, r_1(n)\tilde{r}(n), r_1(n)^2, r_1(n)\frac{\log^{0.25}(K(n))}{(np)^{0.25}}, r_1(n)^{1.5}\tilde{r}(n)^{0.5}\right).$$

If $\frac{r_1(n)}{\tilde{r}(n)} \to 0$ *for* $n \to \infty$*, then*

$$O(H(n)) = O\left(\frac{\log^{0.5}(K(n))}{(np)^{0.5}}\right) + o\left(\tilde{r}(n)\max\left(\tilde{r}(n), \frac{\log^{0.25}(K(n))}{(np)^{0.25}}\right)\right).$$

Thus, if also $\frac{\log(K(n))}{(np)^{0.5}\tilde{r}(n)^2} \to 0$ *for* $n \to \infty$*, then*

$$H(n) = o(\tilde{r}(n)) = o(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}).$$

Asymptotic Optimality. *Thus, if* $\frac{r_1(n)}{\tilde{r}(n)} \to 0$ *and* $\frac{\log^{0.5}(K(n))}{(np)^{0.5}\tilde{r}(n)^2} \to 0$ *for* $n \to \infty$*, then*

$$\frac{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 1 \text{ for } n \to \infty. \tag{36}$$

Finally, if $\frac{r_1(n)}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 0$ *in probability, and* $\frac{\log^{0.5}(K(n)}{(np)^{0.5}\{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\}} \to 0$ *in probability, then*

$$\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 1 \text{ in probability for } n \to \infty. \tag{37}$$

This theorem is a direct corollary of Theorem 2 and Lemma 11.

10 Estimating the risk of the estimator, and confidence intervals.

Let $\hat{\psi} = \psi(\cdot \mid P_n) \equiv \psi_{\hat{k}(P_n)}(\cdot \mid P_n)$ be our selected estimator. Let

$$\tilde{\theta}_n \equiv \int L(O, \psi(\cdot \mid P_n) \mid \eta_0) dP_0(O)$$

be its true conditional risk. Two estimators of this parameter $\tilde{\theta}_n$ can be considered. Firstly, we have the substitution estimator:

$$\hat{\theta} = \int L(O, \psi(\cdot \mid P_n) \mid \eta(P_n)) dP_n(O).$$

Here $\eta(P_n)$ denotes an estimator of η_0. There is typically some concern that this estimator might be biased low. Therefore, the cross-validation estimator might be more accurate in finite samples:

$$\hat{\theta}_{n(1-p^*)} = E_{S_{n*}} \int L(O, \psi(\cdot \mid P^0_{n,S_{n*}}) \mid \eta(P^0_{n,S_{n*}})) dP^1_{n,S_{n*}}(O),$$

where S_{n*} now identifies a split in a learning and test sample and p^* is the proportion of the test sample. Note that evaluating this quantity $\hat{\theta}_{n(1-p)}$ requires double cross-validation in the sense that beyond repeatedly splitting the sample P_n in a test $P^1_{n,S_{n*}}$ and learning sample $P^0_{n,S_{n*}}$, evaluation of the estimator $\psi(\cdot \mid P^0_{n,S_{n*}})$ on a learning sample requires evaluation of our selector $\hat{k}(P^0_{n,S_{n*}})$, which itself requires repeatedly splitting the learning sample in a training and validation sample. Similarly, the estimator $\eta(P_n)$ might itself already involve cross-validation to fine tune parameters.

10.1 Confidence interval for risk.

We will now provide a strategy for constructing a confidence interval for the true conditional risk $\tilde{\theta}_n$. Firstly, we note that $\hat{\theta}_{n(1-p^*)}$ is aiming to estimate the quantaty

$$\tilde{\theta}_{n(1-p^*)} \equiv E_{S_{n*}} \int L(O, \psi(\cdot \mid P^0_{n,S_{n*}}) \mid \eta_0) dP_0(O).$$

Suppose that we can establish a first order linear expansion:

$$\hat{\theta}_{n(1-p^*)} - \tilde{\theta}_{n(1-p^*)} = \frac{1}{n} \sum_{i=1}^{n} IC(O_i \mid P_0) + o_P(1/\sqrt{n})$$

for some fixed function $IC(\cdot \mid P_0)$ of O with mean zero and finite variance $\sigma^2 = \mathrm{VAR}_{P_0} IC(O \mid P_0)$. For example, such a result is proved in Dudoit, van der Laan (2003) for loss functions which do not depend on a nuisance parameter η_0. This asymptotic linearity result allows us to derive confidence intervals for the conditional risk $\tilde{\theta}_{n(1-p^*)}$. From the Central Limit Theorem, as $n \to \infty$, $\sqrt{n}(\hat{\theta}_{n(1-p^*)} - \tilde{\theta}_{n(1-p^*)})/\sigma$ converges in distribution to a standard normal random variable. We can now construct estimators $\hat{IC}(O)$ and $\hat{\sigma}^2 = 1/n \sum_i \hat{IC}(O_i)^2$ of $IC(O \mid P_0)$ and its variance σ^2. An asymptotic $(1-\alpha)100\%$ confidence interval for $\tilde{\theta}_{n(1-p^*)}$ is given by

$$\hat{\theta}_{n(1-p)} \pm z_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}},$$

where $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ for the standard normal cumulative distribution function $\Phi(\cdot)$.

If $\tilde{\theta}_{n(1-p^*)} - \tilde{\theta}_n$ constitutes a second order difference, then this confidence interval also provides an asymptotic $(1 - \alpha)$-confidence interval for $\tilde{\theta}_n$. This method for constructing confidence intervals for a true conditional risk of a given estimator is carried out in detail in Dudoit, van der Laan (2003) in the context of prediction (i.e., Example 1).

The Adaptive epsilon-Net Estimator

11 Introduction.

A general loss function based approach for model selection and estimation, described in Barron et al. (1999), uses sieve theory to define penalized empirical loss criteria. Connections with cross-validation methods are discussed in Birgé and Massart (1997). Birgé and Massart (1997), Barron et al. (1999) have studied thoroughly the penalty functions to be used in the problems of adaptive estimation on sieves. They use powerful Talagrand's concentration and deviation inequalities for empirical processes (Talagrand (1996a), Talagrand (1996b), Ledoux (1996), Massart (1998)) to obtain so called oracle inequalities for the theoretical risk of their estimators. The method of oracle inqualities was also used to prove optimality properties of nonparametric estimators in (Johnstone (1998)). The Birgé-Massart penalties are based on the dimension of the classes of functions. This approach has been shown to perform well in some examples of sieves that frequently occur in nonparametric univariate regression and nonparametric univarite density estimation (nested families of Sobolev ellipsoids).

In this part III of our article we focus on a particular type of sieve, namely ϵ-nets of guessed subspaces (indexed by s) of the complete parameter space, and we use cross-validation as selection criteria. Our results on the cross-validation selector in Part I show that it performs typically asymptotically as well as the oracle selector which for the given data set makes the optimal choice depending on the truth. This shows that cross-validation is a very adaptive procedure and thereby can be expected to be preferable beyond the use of universal (independent of the true distribution) penalty terms. In addition, the sparsity of epsilon-nets makes the epsilon-net a particular effective sieve. Finally, our results apply to loss functions depending on nuisance parameters, so that they can be applied to a very wide range of estimation problems.

Our general estimator is defined as follows. For a collection of subspaces and the complete parameter space, one defines an epsilon-net (i.e., a finite set of points whose ϵ-spheres cover the complete parameter space). For each ϵ and subspace one defines now a corresponding minimum cross-valided empir-

ical risk estimator as the minimizer of cross-validated risk over the subspace-specific ϵ-net. In the special case that the loss function has no nuisance parameter, which thus covers the classical regression and density estimation cases, this ϵ and subspace specific minimum risk estimator reduces to the minimizer of the *empirical risk* over the corresponding ϵ-net (e.g., least squares estimator, maximum likelihood estimator). Finally, one selects ϵ and the subspace with the cross-validation selector. We refer to the resulting estimator as the cross-validated adaptive ϵ-net estimator.

Our estimator can be denoted as $\psi_{\epsilon(P_n),s(P_n)}(P_n)$, where $(\epsilon(P_n), s(P_n))$ equals the minimizer of the cross-validated empirical risk estimate of the estimator $\psi_{\epsilon,s}(P_n)$ over all choices (ϵ, s). We will prove a general finite sample inequality for the marginal risk $Ed(\psi_{\epsilon(P_n),s(P_n)}(P_{n(1-p)}), \psi_0)$, i.e., the marginal risk of the data adaptively selected estimator applied to a subsample of size $n(1-p)$ minus the minimal risk, where p denotes the proportion constituting the validation sample in the employed cross-validation scheme. This finite sample inequality teaches us that the estimator achieves at minimal the minimax rate implied by the size of the parameter space $\mathbf{\Psi}$. However, in addition, by the fact that the estimator chooses data adaptively (using cross-validation) the best choice s and ϵ for the ϵ-net, the finite sample inequality also shows that the estimator is adaptive. Finally, as mentioned above, our theorems on the cross-validation selector teach us that, asymptotically, the cross-validated choice of ϵ and s will typically perform as well as an oracle procedure making the optimal choice (which depends on the unknown P_0) for the given data set.

Le Cam has used ϵ-nets to construct efficient estimators in parametric models, which is often referred to as Le Cam's discretization device (Cam (1986), Cam and Yang (1990)). Our argument in favor of the use of ϵ-nets as sieve is that it yields the smallest (in size, and thereby sparsest) ϵ-approximation of the complete parameter space, while other sieves might yield dense approximations in certain areas of the parameter space, but might result in ineffective approximations at other parts of the parameter space. Recently, Donoho (2003) has argued a theoretical geometrical advantage of ϵ-nets in relation to other choices of sieves in the context of univariate nonparametric regression. This advantage of ϵ-nets are also connected with the sparsity concept as developed in Donoho and Johnstone (1994). We also like to stress that algorithms for nonparametric multivariate regression and conditional density estimation in the statistics and machine learning literature (Hastie et al. (2001),Ripley (1996),Breiman et al. (1984a)) do not corre-

spond with ϵ-net sieves (e.g., many of the sieves correspond with constraints on the norm of the vector of coefficients), and that these algorithms do also not aim to minimize the empirical mean of the loss function (e.g., sum of residual squared errors) over specified subspaces of the complete parameter space. Instead these algorithms rely of forward/backward type of local optimization steps. Our approach provides a road map for developing minimax adaptive estimators in a large class of problems based on ϵ-net sieves and cross-validation.

12 The cross-validated ϵ-net estimator.

Let $\|\cdot\|_{\mathbf{\Psi}}$ be a norm defined on the parameter space $\mathbf{\Psi} = \{\psi(\cdot \mid P) : P \in \mathcal{M}\}$. For each $\epsilon > 0$, let $\mathbf{\Psi}_\epsilon \equiv \{\psi_1^\epsilon, \ldots, \psi_{N(\epsilon)}^\epsilon\} \subset \mathbf{\Psi}$ be a finite set of elements in $\mathbf{\Psi}$ so that the union $\cup_{j=1}^{N(\epsilon)} B(\psi_j^\epsilon, \epsilon)$ of all spheres $B(\psi_j^\epsilon, \epsilon) \equiv \{\psi \in \mathbf{\Psi} : \| \psi - \psi_j^\epsilon \| \leq \epsilon\}$ centered at ψ_j^ϵ with radius ϵ covers the complete parameter space $\mathbf{\Psi}$, $s = 1, \ldots, K_1(n)$. One refers to such a set of elements of $\mathbf{\Psi}$ as an ϵ-net of $\mathbf{\Psi}$. In our case, this finite set functions will be treated as a *discrete sieve*. The minimal number $N(\epsilon, \mathbf{\Psi}, \|\cdot\|_{\mathbf{\Psi}})$ of such balls needed to cover $\mathbf{\Psi}$ is typically refered to as the covering number of $\mathbf{\Psi}$, whose function in ϵ should be viewed as a measure of the size of the parameter space $\mathbf{\Psi}$. Let

$$AE_0(\epsilon) = AE(\epsilon \mid P_0) \equiv \sup_{\{\psi \in \mathbf{\Psi} : \|\psi - \psi_0\|_{\mathbf{\Psi}} \leq \epsilon\}} \int L(O, \psi \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O).$$

denote the with ϵ corresponding worst-case approximation error for the optimal risk. Note that indeed the approximation error for the optimal risk of the discrete sieve $\mathbf{\Psi}_\epsilon$ is bounded by $AE_0(\epsilon)$:

$$B_0(\epsilon) = B(\epsilon \mid P_0) \equiv \min_{\psi \in \mathbf{\Psi}_\epsilon} \int L(O, \psi \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \leq AE_0(\epsilon).$$

Given ϵ, the corresponding discrete sieve $\mathbf{\Psi}_\epsilon$, and a random vector $S_n^* \in \{0,1\}^n$ defining a cross-validation scheme, we define the following estimator:

$$\psi_\epsilon(\cdot \mid P_n) \equiv \text{argmin}_{\psi \in \mathbf{\Psi}_\epsilon} E_{S_{n*}} \int L(O, \psi \mid \eta^0_{n,S_{n*}}) dP^1_{n,S_{n*}}(O).$$

In other words, $\psi_\epsilon(\cdot \mid P_n)$ is the minimizer over the set $\mathbf{\Psi}_\epsilon \subset \mathbf{\Psi}$ of the cross-validated risk. We note that, if the nuisance parameter η_0 is known

(i.e., $\eta^0_{n,S_{n*}} = \eta_0$), then $\psi_\epsilon(\cdot \mid P_n)$ reduces to the minimizer of the empirical average loss:

$$\psi_\epsilon(\cdot \mid P_n) \equiv \text{argmin}_{\psi \in \mathbf{\Psi}_\epsilon} \int L(O, \psi \mid \eta_0) dP_n(O),$$

which does thus not require cross-validation. Let $p^* \equiv \sum_{i=1}^n S_n^*(i)/n$ denote the proportion of the validation sample, where $p^* = 1$ in the latter case with η_0 being known.

We note that for each fixed ϵ, $\psi_\epsilon(\cdot \mid P_n)$ can be viewed as an estimator of the best choice

$$\psi_\epsilon(\cdot \mid P_0) \equiv \text{argmin}_{\psi \in \mathbf{\Psi}_\epsilon} \int L(O, \psi \mid \eta_0) dP(O).$$

For quadratic loss functions, we define

$$\epsilon_{opt}(n) \equiv \text{argmin}_\epsilon B_0(\epsilon) + \frac{\log(N(\epsilon_n))}{n}.$$

In Section 14 it is shown that the estimators $\psi_{c*\epsilon_{opt}(n)}$, indexed by a constant c, converges to the parameter of interest ψ_0 at an optimal (but non-adaptive to properties of ψ_0 itself) rate:

$$E d_n(\hat{\psi}_{c*\epsilon_{opt}(n)}, \psi_0) = O(r_{opt}(n)).$$

A similar statement holds for general loss functions.

In order to obtain an estimator which does not only achieve the optimal rate of convergence, but also chooses the optimal (smoothing) constant c and adapts to the underlying smoothness of ψ_0, we will select c with cross-validation. Let $\epsilon_n(k) = c_k \epsilon_{opt}(n)$, where $c_1 < \ldots < c_{K(n)}$ denotes a partition of an interval $(0, \infty)$. In order to allow the estimator to be adaptive to smoothness of ψ_0 (more than the smoothness assumed in $\mathbf{\Psi}$), one should let $c_{K(n)}$ converge to infinity with sample size n. In other words, the parametrization $\epsilon_n(k) = c_k \epsilon_{opt}(n)$ is used to work at a scale which is already known to be optimal, but we should not enforce this scale in order to allow the estimator to become adaptive to ψ_0.

This discretization $\{\epsilon_n(k) : k = 1, \ldots, K(n)\}$ of a set of possible $\epsilon's$ defines now a sequence of estimators

$$\psi_k(\cdot \mid P_n) \equiv \psi_{\epsilon_n(k)}(\cdot \mid P_n), \ k = 1, \ldots, K(n).$$

We will select k (and thus the choice ϵ) with cross-validation:

$$\hat{k} = \text{argmin}_{k \in \{1,\ldots,K(n)\}} E_{S_n} \int L(O, \psi_{\epsilon_n(k)}(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) dP^1_{n,S_n}(O).$$

Our proposed cross-validated discrete sieve estimator of ψ_0 is given by

$$\psi_{\hat{k}}(\cdot \mid P_n) = \psi_{\epsilon_n(\hat{k})}(\cdot \mid P_n). \tag{38}$$

12.1 Construction of an ϵ-net of the parameter space.

Given the parameter space $(\mathbf{\Psi}, \| \cdot \|_{\mathbf{\Psi}})$ for our parameter of interest $\psi_0 : \mathcal{S} \to \mathbb{R}$, let $\{\phi_j : j\}$ be a countable collection of basis functions so that each $\psi \in \mathbf{\Psi}$ can be arbitrarily well approximated by a finite linear combination of such functions. That is, for each $\psi \in \mathbf{\Psi}$, there exists a countable index set I and a corresponding vector of coefficients $(\beta_j : j \in I)$ so that

$$\psi = \sum_{j \in I} \beta_j \phi_j,$$

where the limit holds w.r.t. the norm $\| \cdot \|_{\mathbf{\Psi}}$. Thus

$$\mathbf{\Psi} = \left\{ \sum_{j \in I} \beta_j \phi_j : \beta \in B_I \subset \mathbb{R}^{|I|} \right\},$$

where the euclidean set B_I is chosen so that $\sum_{j \in I} \beta_j \phi_j \in \mathbf{\Psi}$ for each $\beta \in B_I$. Consequently, by defining

$$\psi_{I,\beta} \equiv \sum_{j \in I} \beta_j \phi_j$$

we have the following parametrization of the parameter space

$$\mathbf{\Psi} = \{\psi_{I,\beta} : \beta \in B_I\}.$$

For a $\delta > 0$, we define

$$B_I(\delta) = \{\beta \in B_I : \forall j, \beta_j/\delta \text{ is an integer}\}.$$

Now, we can define

$$\mathbf{\Psi}_\delta = \{\psi_{I,\beta} : \beta \in B_I(\delta)\}.$$

Note that $\mathbf{\Psi}_\delta$ is finite, and for each $\epsilon > 0$, there exist a $\delta(\epsilon)$ so that $\mathbf{\Psi}_{\delta(\epsilon)}$ is an ϵ-net of $\mathbf{\Psi}$.

Alternatively, we can apply this method for constructing an ϵ-net to a continuous sieve approximation of $\mathbf{\Psi}$, as follows. Given an $\epsilon > 0$, let $I_\epsilon \in \mathcal{I}$ be a finite index set of size $N_1(\epsilon)$ so that the corresponding finite set $\{\phi_j, j \in I_\epsilon\}$ of basis functions generates an ϵ-approximation of $\mathbf{\Psi}$. That is, for each $\psi \in \mathbf{\Psi}$, we have

$$\inf_{\beta \in B_{I_\epsilon}} \parallel \psi - \sum_{j \in I_\epsilon} \beta_j \phi_j \parallel_{\mathbf{\Psi}} \leq \epsilon.$$

Let

$$\mathbf{\Psi}_\epsilon \equiv \{\psi_{I_\epsilon,\beta} : \beta \in B_{I_\epsilon} \subset \mathbb{R}^{N_1(\epsilon)}\}$$

be the corresponding element of the continuous sieve $(\mathbf{\Psi}_\epsilon : \epsilon)$ indexed by ϵ. For a $\delta > 0$, we define

$$B_{I_\epsilon}(\delta) = \{\beta \in B_{I_\epsilon} : \forall j, \beta_j/\delta \text{ is an integer}\}.$$

Now, we can define

$$\mathbf{\Psi}_{\delta,\epsilon} = \{\psi_{I_\epsilon,\beta} : \beta \in B_{I_\epsilon}(\delta)\}.$$

Note that $\mathbf{\Psi}_{\delta,\epsilon}$ is finite, and for each $\epsilon > 0$, there exist a $\delta(\epsilon)$ so that $\mathbf{\Psi}_{\delta(\epsilon),\epsilon}$ is an $2 * \epsilon$-net of $\mathbf{\Psi}$.

Orthonormalizing basis functions. The delta-nets are equally spaced sets of points in the euclidean space for β, and do therefore not necessarily result in sensible (i.e., equally spaced) ϵ-nets in the actual parameter space $\mathbf{\Psi}^s$. Therefore, to construct more equally spaced ϵ-nets (and thus sparser ϵ-nets) one should first orthonormalize the basis functions (in case, the basis we start out with is not orthonormal), and construct the δ-nets based on the corresponding orthonormal parametrization of $\mathbf{\Psi}^*$.

12.2 Algorithm for minimizing over an ϵ-net.

Consider a function $f : \mathbf{\Psi}_\delta \to \mathbb{R}$. Suppose that we wish to minimize this function. Firstly, we note that each $\beta \in B_I(\delta)$ can be identified by an element in the lattice $\{-M, -(M-1), \ldots, 0, 1, \ldots, M\}^{|I|}$, where M is a finite integer. Let $f^* : \{-M, -(M-1), \ldots, 0, 1, \ldots, M\}^{|I|} \to \mathbb{R}$ be an extension of f to this lattice, where we define $f^*(x) = \infty$ for any x which does not correspond with a point in $B_I(\delta)$. We propose the following simple algorithm for minimizing such a function f^* on a lattice $\{0, 1, \ldots, M\}^{|I|}$.

Inititialize Set $k = 0$, and $x_k = (0, 0, \ldots, 0)$.

Define moves For any $x \in \{0, 1, \ldots, M\}^{|I|}$, let $\mathcal{S}(x)$ be defined as the set of $2 * \mid I \mid$ vectors one obtains by adding 1 or subtracting 1 from a particular component x_j. In the case that such moves result in parameters outside the parameter space $\mathbf{\Psi}_s$, then one should augment this set of moves to guarantee a rich enough set of moves (e.g., if adding 1 results in a parameter outside the parameter space, then one can set any of the non-zero components equal to zero).

Iterate Let

$$x^* = \text{argmin}_{x \in \mathcal{S}(x_k)} f^*(x).$$

If $f(x^*) \leq f(x_k)$, then $k = k + 1$, $x_k = x^*$. Otherwise, stop.

Output Let the final x^* be the candidate for the global minimum $j(I \mid P_n)$.

Starting values One could run this algorithm with various starting values.

13 The adaptive cross-validated ϵ-net estimator.

Let $\mathbf{\Psi}_s \subset \mathbf{\Psi}$, $s = 1, \ldots, K_1(n)$, be a collection of subspaces. For each s and each $\epsilon > 0$, let $\mathbf{\Psi}_{\epsilon,s} \equiv \{\psi_1^{\epsilon,s}, \ldots, \psi_{N_s(\epsilon)}^{\epsilon,s}\} \subset \mathbf{\Psi}_s$ be a finite set of elements in $\mathbf{\Psi}_s$ so that the union $\cup_{j=1}^{N_s(\epsilon)} B(\psi_j^{\epsilon,s}, \epsilon)$ of all spheres $B(\psi_j^{\epsilon,s}, \epsilon) \equiv \{\psi \in \mathbf{\Psi}_s : \| \psi - \psi_j^{\epsilon,s} \| \leq \epsilon\}$ centered at $\psi_j^{\epsilon,s}$ with radius ϵ covers the complete parameter space $\mathbf{\Psi}_s$, $s = 1, \ldots, K_1(n)$. One refers to such a set of elements of $\mathbf{\Psi}_s$ as an ϵ-net of $\mathbf{\Psi}_s$, $s = 1, \ldots, K_1(n)$. In our case, this finite set functions will be treated as a *discrete sieve*. The minimal number $N_s(\epsilon, \mathbf{\Psi}_s, \| \cdot \|_{\mathbf{\Psi}}$ of such balls needed to cover $\mathbf{\Psi}_s$ is typically refered to as the covering number of $\mathbf{\Psi}_s$, whose function in ϵ should be viewed as a measure of the size of the parameter space $\mathbf{\Psi}_s$, $s = 1, \ldots, K_1(n)$. Let

$$AE_0(\epsilon, s) = AE(\epsilon, s \mid P_0) \equiv \sup_{\{\psi \in \mathbf{\Psi}_s : \|\psi - \psi_0\|_{\mathbf{\Psi}} \leq \epsilon\}} \int L(O, \psi \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O).$$

denote the with ϵ corresponding worst-case approximation error for the optimal risk. Note that indeed the approximation error for the optimal risk of

the discrete sieve $\mathbf{\Psi}_{\epsilon,s}$ is bounded by $AE_0(\epsilon, s)$:

$$B_0(\epsilon, s) = B(\epsilon, s \mid P_0) \equiv \min_{\psi \in \mathbf{\Psi}_{\epsilon,s}} \int L(O, \psi \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \leq AE_0(\epsilon, s).$$

Given (ϵ, s), the corresponding discrete sieve $\mathbf{\Psi}_{\epsilon,s}$, and a random vector $S_n^* \in \{0,1\}^n$ defining a cross-validation scheme, we define the following estimator:

$$\psi_{\epsilon,s}(\cdot \mid P_n) \equiv \text{argmin}_{\psi \in \mathbf{\Psi}_{\epsilon,s}} E_{S_{n*}} \int L(O, \psi \mid \eta^0_{n,S_{n*}}) dP^1_{n,S_{n*}}(O).$$

In other words, $\psi_{\epsilon,s}(\cdot \mid P_n)$ is the minimizer over the set $\mathbf{\Psi}_{\epsilon,s} \subset \mathbf{\Psi}_s$ of the cross-validated risk. We note that, if the nuisance parameter η_0 is known (i.e., $\eta^0_{n,S_{n*}} = \eta_0$), then $\psi_{\epsilon,s}(\cdot \mid P_n)$ reduces to the minimizer of the empirical average loss:

$$\psi_{\epsilon,s}(\cdot \mid P_n) \equiv \text{argmin}_{\psi \in \mathbf{\Psi}_{\epsilon,s}} \int L(O, \psi \mid \eta_0) dP_n(O),$$

which does thus not require cross-validation. Let $p^* \equiv \sum_{i=1}^n S_n^*(i)/n$ denote the proportion of the validation sample, where $p^* = 1$ in the latter case with η_0 being known.

We note that for each fixed (s, ϵ), $\psi_{\epsilon,s}(\cdot \mid P_n)$ can be viewed as an estimator of the best choice

$$\psi_{\epsilon,s}(\cdot \mid P_0) \equiv \text{argmin}_{\psi \in \mathbf{\Psi}_{\epsilon,s}} \int L(O, \psi \mid \eta_0) dP(O).$$

For each s, let $\epsilon_n(k)$, $k = 1, \ldots, K_{2s}(n)$ be a given set of values. Let

$$\mathcal{A}_n = \cup_{s=1}^{K_1(n)} \{(s, \epsilon_n(k, s) : k \in \{1, \ldots, K_{2s}(n)\}.$$

Let

$$K(n) = \mid \mathcal{A}_n \mid \text{ be the size of } \mathcal{A}_n.$$

We select (s, ϵ) with cross-validation:

$$(\hat{s}(P_n), \hat{\epsilon}(P_n)) \equiv \text{argmin}_{(s,\epsilon) \in \mathcal{A}_n} E_{S_n} \int L(O, \psi_{s,\epsilon}(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) dP^1_{n,S_n}(O).$$

The adaptive cross-validated discrete sieve estimator is now defined by

$$\hat{\psi}(\cdot \mid P_n) = \psi_{\hat{s}(P_n), \hat{\epsilon}(P_n)}(\cdot \mid P_n).$$

In Section 16 we prove that this estimator achieves the optimal rate of convergence corresponding with the smallest subspace $\mathbf{\Psi}_{s^*}$, as measured by the covering numbers $N_s(\epsilon)$, which contains ψ_0.

14 Finite sample results for epsilon-net estimator

14.1 Quadratic loss function.

The next theorem provides us with a finite sample bound for the difference of the conditional risks of the discrete sieve estimator $\psi_\epsilon(\cdot \mid P_n)$ and $\psi_\epsilon(\cdot \mid P_0)$, for a given sequence ϵ_n. This finite sample bound provides, in particular, an optimal rate $\epsilon_{opt}(n)$ at which ϵ should converge to zero with sample size n. This theorem follows from a direct application of our general Theorems 1.

Theorem 5 *(***Finite sample result and Asymptotics for** $\psi_{\epsilon_n}(\cdot \mid P_n)$*)* *Let $\epsilon_n > 0$ be a given sequence converging to zero with sample size n.*
Assumptions.
A1. *The limit η_0 of $\eta_n = \eta(P_n)$ for $n \to \infty$ is an element of $\Gamma(P_0)$.*
A2. *There exist a $M_1^* < \infty$ so that*

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{O} L^*(O, \psi, \psi_0) \leq M_1^*,$$

where the supremum is taken over a support of the distribution P_0 of O.
A3. *There exist a $M_2 < \infty$ so that*

$$\sup_{\psi \in \mathbf{\Psi}} \frac{VAR_{P_0} L^*(O, \psi, \psi_0)}{E_{P_0} L^*(O, \psi, \psi_0)} \leq M_2.$$

Definitions. *We define the following constants:*

$$\begin{aligned} M_1 &= 2M_1^* \\ c(M_1, M_2, \delta) &= 2(1+\delta)^2 \left(\frac{M_1}{3} + \frac{M_2}{\delta} \right) \\ a_0 &\equiv 2M_1/3 \\ M_3(N(\epsilon_n)) &= 3a_0 + \sqrt{2}\frac{\sqrt{\log(2)}}{\log(N(\epsilon_n))} + \frac{\sqrt{2}}{\sqrt{\log(N(\epsilon_n))}} + b_0 + \int_{b_0}^{\infty} 2N(\epsilon_n)^{1-m(x)} dx, \end{aligned}$$

where b_0 is the smallest constant larger than the solution of $1 - m(x) = 0$ with $m(x) = 0.5\frac{x^2}{1/\log(N(\epsilon_n)) + a_0 x}$. We note that $M_3(\log(N(\epsilon_n)) \downarrow$ in n. Let $\hat{k}$ and $\tilde{k}$ be defined by $\psi_{\epsilon_n}(\cdot \mid P_n) = \psi_{\hat{k}}^{\epsilon_n}$ and $\psi_{\epsilon_n}(\cdot \mid P_0) = \psi_{\tilde{k}}^{\epsilon_n}$ We also define

the following sequences in n:

$$r_1^*(n) \equiv \max_{\bar{k}\in\{\hat{k},\tilde{k}\}} \frac{E\int(L_{n,S_{n*}}^{*0}-L^*)(O,\psi_{\bar{k}}^{\epsilon_n},\psi_0)dP_0(O)}{\sqrt{E\int L^*(O,\psi_{\bar{k}}^{\epsilon_n},\psi_0)dP_0(O)}}$$

$$r_2^*(n) \equiv E\max_{k\in\{1,\ldots,N(\epsilon_n)\}}\sqrt{\int(L_{n,S_{n*}}^{*0}-L^*)^2(O,\psi_k^{\epsilon_n},\psi_0)dP_0(O)}$$

Finally, for any $\delta>0$ we define

$$\epsilon_n(\delta) \equiv (1+2\delta)B_0(\epsilon_n)+2c(M_1,M_2,\delta)\frac{1+\log(N(\epsilon_n))}{np^*}$$
$$+(1+\delta)r_1^*(n)\sqrt{B(\epsilon_n)}+\frac{2M_3(N(\epsilon_n))(1+\delta)\log(N(\epsilon_n))}{(np^*)^{0.5}}\max(r_2^*(n),(np^*)^{-0.5}I(r_2^*(n)>0)).$$

Finite Sample Result. *For any $\delta>0$, we have*

$$Ed_n(\psi_{\epsilon_n}(\cdot\mid P_n),\psi_0) \leq \left\{\frac{r_1^*(n)(1+\delta)+\sqrt{r_1^*(n)^2(1+\delta)^2+4\epsilon_n(\delta)}}{2}\right\}^2$$
$$\equiv f_1(B_0(\epsilon_n),r_1^*(n),r_2^*(n),\log(N(\epsilon_n)),np^*,M_1,M_2,M_3(N(\epsilon_n)),\delta).$$

If $\eta_n=\eta_0$ is known, then this reduces to (we can set $p^*=1$)

$$Ed_n(\psi_{\epsilon_n}(\cdot\mid P_n),\psi_0)\leq(1+2\delta)B_0(\epsilon_n)+2c(M_1,M_2,\delta)\frac{1+\log(N(\epsilon_n))}{n}.$$

Asymptotic Implication. *For any $\delta>0$*

$$Ed_n(\psi_{\epsilon_n}(\cdot\mid P_n),\psi_0) \leq (1+2\delta)B_0(\epsilon_n)+O(H(n)),$$

where

$$H(n) \equiv \max\left(\frac{\log(N(\epsilon_n))}{np^*},\frac{\log(N(\epsilon_n))r_2^*(n)}{(np^*)^{0.5}},r_1^*(n)^2,r_1^*(n)\sqrt{B_0(\epsilon_n)},r_1^*(n)^{1.5}B_0(\epsilon_n)^{0.25},\right.$$
$$\left.\frac{\sqrt{\log(N(\epsilon_n))}r_1^*(n)}{(np^*)^{0.5}},\frac{\sqrt{\log(N(\epsilon_n))}r_2^*(n)^{0.5}r_1^*(n)}{(np^*)^{0.25}}\right).$$

Consequently, we have the following scenarios.
Optimal rate: *If* $\max\left(\frac{\log(N(\epsilon_n))}{np^*},r_1^*(n)^2,\log(N(\epsilon_n))r_2^*(n)^2\right)=O(B_0(\epsilon_n))$, *then* $H(n)=O(B_0(\epsilon_n))$, *and thus*

$$Ed_n(\psi_{\epsilon_n}(\cdot\mid P_n),\psi_0)=O(B_0(\epsilon_n)).$$

Define

$$\begin{aligned} r_{opt}(n) &\equiv \min_{\epsilon} B_0(\epsilon) + \frac{\log(N(\epsilon_n))}{n} \\ \epsilon_{opt}(n) &\equiv argmin_{\epsilon} \left\{ B_0(\epsilon) + \frac{\log(N(\epsilon_n))}{n} \right\}. \end{aligned}$$

It follows that, if $\max\left(r_1^*(n)^2, \log(N(\epsilon_n))r_2^*(n)^2\right) = O(B_0(\epsilon_n))$ *and* $\epsilon_n = c * \epsilon_{opt}(n)$ *for a* $0 < c < \infty$*, then*

$$\begin{aligned} Ed_n(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0) &= E \int L(O, \psi_{\epsilon_n}(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\ &= O(B_0(\epsilon_n)) = O(r_{opt}(n)). \end{aligned}$$

Asymptotic equivalence with best choice $\psi_{\epsilon_n}(\cdot \mid P_0)$**:**
If either $\max\left(\frac{\log(N(\epsilon_n))}{np^*}, r_1^*(n)^2, \log(N(\epsilon_n))r_2^*(n)^2\right) = o(B_0(\epsilon_n))$ *or* $\max\left(\frac{\log(N(\epsilon_n))^2}{np^*}, r_1^*(n)^2, r_2^*(n)^2\right) = o(B_0(\epsilon_n))$*, then*

$$H(n) = o(B_0(\epsilon_n)).$$

In particular, we note that if $\max(r_1^*(n)^2, \log(N(\epsilon_n))r_2^*(n)^2) = o(B_0(\epsilon_n))$*, then*

$$H(n) = O\left(\frac{\log(N(\epsilon_n))}{np^*}\right) + o(B_0(\epsilon_n)).$$

Asymptotic Optimality. *Consequently, under these two possible scenarios for which* $H(n) = o(B_0(\epsilon_n))$*, we have*

$$\frac{Ed_n(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0)}{Ed_n(\psi_{\epsilon_n}(\cdot \mid P_0), \psi_0)} \to 1 \text{ for } n \to \infty, \tag{39}$$

and, in particular,

$$\frac{d_n(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0)}{d_n(\psi_{\epsilon_n}(\cdot \mid P_0), \psi_0)} \to 1 \text{ in probability for } n \to \infty.$$

Proof. We apply Theorem 1 with candidate "estimators" $\psi_k(\cdot \mid P_n) \equiv \psi_k^{\epsilon_n}$ (constants), $k = 1, \ldots, N(\epsilon_n)$. Note that in this setting the quantaties in Theorem 1 have the following corresponding analogues:

$$\psi_k(\cdot \mid P_n) = \psi_k^{\epsilon_n}$$

$$
\begin{aligned}
\psi_{\hat{k}}(\cdot \mid P_n) &= \psi_{\epsilon_n}(\cdot \mid P_n) \\
\psi_{\tilde{k}}(\cdot \mid P_n) &= \text{argmin}_{\psi \in \mathbf{\Psi}_{\epsilon_n}} \int L(O, \psi \mid \eta_0) dP_0(O) \\
&\equiv \psi_{\epsilon_n}(\cdot \mid P_0). \\
L^*(O, \psi_k(\cdot \mid P^*_{n,S_n}), \psi_0) &= L(O, \psi_k^{\epsilon_n} \mid \eta_0) - L(O, \psi_0 \mid \eta_0) \\
d_{n(1-p^*)}(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0) &= \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \\
&= d_n(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0) \\
&= \int L(O, \psi_{\epsilon_n}(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\
d_{n(1-p^*)}(\psi_{\epsilon_n}(\cdot \mid P_0), \psi_0) &= \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} \\
&= d_n(\psi_{\epsilon_n}(\cdot \mid P_0), \psi_0) \\
&= \int L(O, \psi_{\epsilon_n}(\cdot \mid P_0) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\
&= B_0(\epsilon_n) \\
K(n) &= N(\epsilon_n) \\
\tilde{r}^2(n) &= B_0(\epsilon_n) \\
&= \int L(O, \psi_{\epsilon_n}(\cdot \mid P_0) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O).
\end{aligned}
$$

The above theorem is now the complete analogue of our general Theorem 1. □

14.2 General loss function.

Theorem 6 *(***Finite sample result and Asymptotics for** $\psi_{\epsilon_n}(\cdot \mid P_n)$*)* *Let* $\epsilon_n > 0$ *be a given sequence converging to zero with sample size* n.
Assumptions.
A1. *The limit* η_0 *of* $\eta^0_{n,S_{n*}}$ *for* $n \to \infty$ *needs to be an element of* $\Gamma(P_0)$.
A2. *There exist a* $M_1^* < \infty$ *so that*

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{O} L^*(O, \psi, \psi_0) \leq M_1^*,$$

where the supremum is taken over the support of the distribution P_0 *of* O.

Definitions. *Let* $\hat{k}, \tilde{k}$ *be defined by* $\psi_{\epsilon_n}(\cdot \mid P_n) = \psi_{\hat{k}}^{\epsilon_n}$ *and* $\psi_{\epsilon_n}(\cdot \mid P_0) = \psi_{\tilde{k}}^{\epsilon_n}$. *We define the following sequences in* n*:*

$$f(M_1^*, N(\epsilon_n), np^*) \equiv 4M_1^{*2}\sqrt{\frac{\log K(n)}{np}}$$

$$
\begin{aligned}
r_1^*(n) &\equiv \max_{\tilde{k}\in\{\hat{k},\tilde{k}\}} \frac{E\int (L^{*0}_{n,S_{n*}} - L^*)(O,\psi_{\tilde{k}}^{\epsilon_n},\psi_0)dP_0(O)}{\sqrt{E\int L^*(O,\psi_{\tilde{k}}^{\epsilon_n},\psi_0)dP_0(O)}} \\
\tilde{r}(n) &\equiv \sqrt{B_0(\epsilon_n)}.
\end{aligned}
$$

We also define

$$
\epsilon_{1n} \equiv B_0(\epsilon_n) + f(M_1^*, N(\epsilon_n), np^*) + r_1^*(n)\sqrt{B_0(\epsilon_n)}.
$$

Finite Sample Result. *We have*

$$
\begin{aligned}
Ed_n(\psi_{\epsilon_n}(\cdot \mid P_n),\psi_0) &\leq \left\{\frac{r_1^*(n) + \sqrt{r_1^*(n)^2 + 4\epsilon_{1n}}}{2}\right\}^2 \\
&= f_2(B_0(\epsilon_n), r_1^*(n), \log(N(\epsilon_n)), np^*, M_1^*).
\end{aligned}
$$

If $\eta_n = \eta_0$ is known, then this reduces to

$$
Ed_n(\psi_{\epsilon_n}(\cdot \mid P_n),\psi_0) \leq B_0(\epsilon_n) + f(M_1^*, N(\epsilon_n), n).
$$

Asymptotic Implication. *We have*

$$
Ed_n(\psi_{\epsilon_n}(\cdot \mid P_n),\psi_0) \leq Ed_n(\psi_{\epsilon_n}(\cdot \mid P_0),\psi_0) + O(H(n)),
$$

where

$$
H(n) \equiv \max\left(\frac{\log^{0.5}(N(\epsilon_n))}{\sqrt{np^*}}, r_1^*(n)\sqrt{B_0(\epsilon_n)}, r_1^*(n)^2, r_1^*(n)\frac{\log^{0.25}(N(\epsilon_n)}{(np^*)^{0.25}}, r_1^*(n)^{1.5}B(\epsilon_n)^{0.25}\right).
$$

Optimal rate: *If* $\max\left(r_1^*(n)^2, \frac{\log^{0.5}(N(\epsilon_n))}{(np^*)^{0.5}}\right) = O(B_0(\epsilon_n))$, *then* $H(n) = O(B_0(\epsilon_n))$, *and thus*

$$
Ed_n(\psi_{\epsilon_n}(\cdot \mid P_n),\psi_0) = O(B_0(\epsilon_n)).
$$

Define

$$
\begin{aligned}
r_{opt}(n) &\equiv \min_{\epsilon}\left\{B_0(\epsilon) + \frac{\log^{0.5}(N(\epsilon))}{\sqrt{n}}\right\} \\
\epsilon_{opt}(n) &\equiv argmin_{\epsilon}\left\{B_0(\epsilon) + \frac{\log^{0.5}(N(\epsilon))}{n^{0.5}}\right\}.
\end{aligned}
$$

It follows that, if $r_1^*(n)^2 = O(B_0(\epsilon_n))$ *and* $\epsilon_n = c * \epsilon_{opt}(n)$ *for a* $0 < c < \infty$, *then*

$$E \int L(O, \psi_{\epsilon_n}(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) = O(B_0(\epsilon_n)) = O(r_{opt}(n)).$$

Asymptotic Equivalence. *If* $r_1^*(n)^2 = o(B_0(\epsilon_n)$, *then*

$$O(H(n)) \quad = \quad O\left(\frac{\log^{0.5}(N(\epsilon_n))}{(np^*)^{0.5}}\right) + o(B_0(\epsilon_n)).$$

Thus, if also $\frac{\log^{0.5}(N(\epsilon_n))}{(np^*)^{0.5}} = o(B_0(\epsilon_n))$, *then*

$$H(n) = o(B_0(\epsilon_n))$$

Consequently, if $\max\left(r_1^*(n)^2, \frac{\log^{0.5}(N(\epsilon_n)}{(np^*)^{0.5}}\right) = o(B_0(\epsilon_n)^2)$, *then*

$$\frac{Ed_n(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0)}{Ed_n(\psi_{\epsilon_n}(\cdot \mid P_0), \psi_0)} \to 1 \text{ for } n \to \infty,$$

and, in particular,

$$\frac{d_n(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0)}{d_n(\psi_{\epsilon_n}(\cdot \mid P_0), \psi_0)} \to 1 \text{ in probability for } n \to \infty.$$

Proof. We apply Theorem 2 with candidate "estimators" $\psi_k(\cdot \mid P_n) \equiv \psi_k^{\epsilon_n}$ (constants), $k = 1, \ldots, N(\epsilon_n)$. Note that in this setting the quantaties in Theorem 2 have the following corresponding analogues:

$$\begin{aligned}
\psi_k(\cdot \mid P_n) &= \psi_k^{\epsilon_n} \\
\psi_{\hat{k}}(\cdot \mid P_n) &= \psi_{\epsilon_n}(\cdot \mid P_n) \\
\psi_{\tilde{k}}(\cdot \mid P_n) &= \text{argmin}_{\psi \in \mathbf{\Psi}_{\epsilon_n}} \int L(O, \psi \mid \eta_0) dP_0(O) \\
&\equiv \psi_{\epsilon_n}(\cdot \mid P_0). \\
L^*(O, \psi_k(\cdot \mid P^*_{n,S_n}), \psi_0) &= L(O, \psi_k^{\epsilon_n} \mid \eta_0) - L(O, \psi_0 \mid \eta_0) \\
d_{n(1-p^*)}(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0) &= \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \\
&= d_n(\psi_{\epsilon_n}(\cdot \mid P_n), \psi_0) \\
&= \int L(O, \psi_{\epsilon_n}(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)
\end{aligned}$$

$$
\begin{aligned}
d_{n(1-p^*)}(\psi_{\epsilon_n}(\cdot \mid P_0), \psi_0) &= \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} \\
&= d_n(\psi_{\epsilon_n}(\cdot \mid P_0), \psi_0) \\
&= \int L(O, \psi_{\epsilon_n}(\cdot \mid P_0) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\
&= B_0(\epsilon_n) \\
K(n) &= N(\epsilon_n) \\
\tilde{r}^2(n) &= B_0(\epsilon_n) \\
&= \int L(O, \psi_{\epsilon_n}(\cdot \mid P_0) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O).
\end{aligned}
$$

The above theorem is now the complete analogue of Theorem 2. □

14.3 Asymptotic implications for ϵ-net estimator.

If $N(\epsilon)$ is of the same order as the covering number $N(\epsilon, \mathbf{\Psi}, \|\cdot\|_{\mathbf{\Psi}}))$, then the above theorems shows that for any constant $c > 0$ the risk of the estimator $\psi_{c*\epsilon_{opt}(n)}$ converges to the optimal risk of ψ_0 at a rate as fast or faster than $r^*_{opt}(n)$, where $r^*_{opt}(n)$ is an explicitly known rate defined as follows. For loss functions satisfying Assumption A3, we have

$$
r_{opt}(n) \leq r^*_{opt}(n) = \min_{\epsilon} \left\{ AE_0(\epsilon) + \frac{\log(N(\epsilon, \mathbf{\Psi}, \|\cdot\|_{\mathbf{\Psi}}))}{n} \right\},
$$

while for loss functions not satisfying Assumption A3, we have

$$
r_{opt}(n) \leq r^*_{opt}(n) = \min_{\epsilon} \left\{ AE_0(\epsilon) + \frac{\log^{0.5}(N(\epsilon, \mathbf{\Psi}, \|\cdot\|_{\mathbf{\Psi}}))}{\sqrt{n}} \right\}.
$$

Let $\epsilon^*_{opt}(n)$ be the argument of the minimum corresponding with $r^*_{opt}(n)$. For quadratic loss functions satisfying Assumption A3 of Theorem 1 we have $AE_0(\epsilon) \leq C\epsilon^2$ for some $C < \infty$. Thus, in this case the rate of convergence $r^*_{opt}(n)$ can be bounded as follows:

$$
r^*_{opt}(n) \leq \min_{\epsilon} \epsilon^2 + \frac{\log(N(\epsilon, \mathbf{\Psi}, \|\cdot\|_{\mathbf{\Psi}}))}{n}.
$$

For non-quadratic loss functions not satisfying Assumption A3, we typically have $AE_0(\epsilon) \leq C\epsilon$ for some $C < \infty$, so that the rate of convergence can be bounded as follows:

$$
r^*_{opt}(n) \leq \min_{\epsilon} \epsilon + \frac{\log^{0.5}(N(\epsilon, \mathbf{\Psi}, \|\cdot\|_{\mathbf{\Psi}}))}{\sqrt{n}}.
$$

Below we will verify that for the well known smoothness classes $\mathbf{\Psi}$ for multivariate real valued functions, the above explicit bounds for $r_{opt}(n)$ and $r^*_{opt}(n)$ correspond with the optimal rates of convergence given in the literature. We refer to van der Vaart and Wellner (1996) for the covering numbers $N(\epsilon, \mathbf{\Psi}, \| \cdot \|_{\mathbf{\Psi}})$ of a variety of classes of functions $\mathbf{\Psi}$.

Since $B_0(\epsilon)$ and $AE_0(\epsilon)$ depend on underlying smoothness of ψ_0, the rate of convergence $r_{opt}(n)$ reported in our theorems could be significantly better than the above bounds $r^*_{opt}(n)$. In other words, the estimator $\psi_{\epsilon_{opt}(n)}(\cdot \mid P_n)$ is capable of adapting to the actual smoothness of ψ_0 and thereby possibly achieves a better rate of convergence than the optimal rate implied by the size of the parameter space $\mathbf{\Psi}$.

14.4 Examples of covering numbers.

Results on covering numbers $N(\epsilon, \mathbf{\Psi}, \| \cdot \|)$ w.r.t. to the supremum norm or other norms can be found in approximation theory. We refer to van der Vaart, Wellner (1996, section 2.7) for a number of very general examples of classes of functions $\mathbf{\Psi}$ and proofs. We will describe here one of their examples.

Example 8 *(***Lipschitz functions on euclidean sets, Theorem 2.7.1, van der Vaart, Wellner, 1996***)* Define for any vector $k = (k_1, \ldots, k_d)$ of d integers the differential operator

$$D^k = \frac{d^{k.}}{dx_1^{k_1} \ldots dx_d^{k_d}},$$

where $k. = \sum_i k_i$. For a number α, let $\underline{\alpha}$ be the largest integer smaller than α. For a function $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$, let

$$\| f \|_\alpha = \max_{k. \leq \alpha} \sup_x \mid D^k f(x) \mid + \max_{k. = \alpha} \sup_{x,y} \frac{\mid D^k f(x) - D^k f(y) \mid}{\| x - y \|^{\alpha - \underline{\alpha}}}$$

Here the suprema are taken over all x, y in the interior of $\mathcal{X}$ with $x \neq y$. Let C^α_M be the set of all continuous functions $f : \mathcal{X} \to \mathbb{R}$ with $\| f \|_\alpha \leq M$. Let $\mathcal{X}$ be a bounded convex subset of $\mathbb{R}^d$ with non-empty interior. There exists a constant K depending only on α and d such that

$$\log(N(\epsilon, C^\alpha_1(\mathcal{X}), \| \cdot \|_\infty) \leq K \lambda(\mathcal{X}_1) \left(\frac{1}{\epsilon}\right)^{d/\alpha}$$

for every $\epsilon > 0$, where $\lambda(\mathcal{X}_1)$ is the Lebesgue measure of the set $\{x : \| x - \mathcal{X} \| < 1\}$. It follows that if $\mathbf{\Psi} = C_M(\alpha)$ for some $M < \infty$, and $AE(\epsilon) \leq \epsilon^2$, then

$$\epsilon^*_{opt}(n) \leq n^{-\frac{\alpha}{2\alpha+d}},$$

which corresponds with a rate $r^*_{opt}(n) = \epsilon^*_{opt}(n)^2 \leq n^{-\frac{2\alpha}{2\alpha+d}}$ which is known to be optimal.

Let $\mathcal{F}$ be the class of all univariate real valued monotone functions uniformly bounded from below and above. Van der Vaart, Wellner (1996) also show that $\log(N(\epsilon, \mathcal{F}, L_r(Q)) \leq \frac{K}{\epsilon}$ for every probability measure Q, every $r \geq 1$ and a constant K only depending on r. Thus, if $\mathbf{\Psi}$ consists of the class of all monotone bounded univariate real valued functions, then $\epsilon_{opt}(n) = n^{-1/3}$ so that $r_{opt}(n) = n^{-2/3}$.

15 Finite sample result for the cross-validated epsilon-net estimator.

Consider the estimator $\psi_{\hat{k}}(\cdot \mid P_n) = \psi_{\epsilon_n(\hat{k})}(\cdot \mid P_n)$ defined in (38). Let $\tilde{k}$ be the optimal comparable benchmark selector for choosing the constant c_k given by

$$\tilde{k} = \text{argmin}_{k \in \{1,\ldots,K(n)\}} E_{S_n} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta_0) dP_0(O)$$

and let

$$\tilde{k}_n = \text{argmin}_{k \in \{1,\ldots,K(n)\}} \int L(O, \psi_k(\cdot \mid P_n) \mid \eta_0) dP_0(O)$$

be the selector for c_k which for each given data set chooses the optimal constant.

15.1 Quadratic loss function.

The following theorem establishes a finite sample bound, and the asymptotic equivalence of $\psi_{\hat{k}}(\cdot \mid P_n)$ with $\psi_{\tilde{k}_n}(\cdot \mid P_n)$ if $K(n)$ is chosen to converge to infinity slowly enough with sample size n so that $\log(K(n))/np$ converges fast enough to zero. This theorem is a direct corollary of Theorem 1 and the asymptotic optimality Corollary 1.

Theorem 7 Assumptions.
A1. *The limit* η_0 *of* $\eta_n = \eta(P_n)$ *for* $n \to \infty$ *is an element of* $\Gamma(P_0)$.
A2. *There exist a* $M_1^* < \infty$ *so that*

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{O} L^*(O, \psi, \psi_0) \leq M_1^*,$$

where the supremum is taken over a support of the distribution P_0 *of* O.
A3. *There exist a* $M_2 < \infty$ *so that*

$$\sup_{\psi \in \mathbf{\Psi}} \frac{\int L^{*2}(O, \psi, \psi_0) dP_0(O)}{\int L^*(O, \psi, \psi_0) dP_0(O)} \leq M_2.$$

Definitions. *We define the following constants:*

$$\begin{aligned}
M_1 &= 2M_1^* \\
c(M_1, M_2, \delta) &= 2(1+\delta)^2 \left(\frac{M_1}{3} + \frac{M_2}{\delta} \right) \\
a_0 &\equiv 2M_1/3 \\
M_3(\log(K(n))) &= 3a_0 + \sqrt{2} \frac{\sqrt{\log(2)}}{\log(K(n))} + \frac{\sqrt{2}}{\sqrt{\log(K(n))}} + b_0 + \int_{b_0}^{\infty} 2K(n)^{1-m(x)} dx,
\end{aligned}$$

where b_0 *is the smallest constant larger than the solution of* $1 - m(x) = 0$ *with* $m(x) = 0.5 \frac{x^2}{1/\log(K(n)) + a_0 x}$. *We note that* $M_3(\log(K(n)) \downarrow$ *in* n.

We also define the following sequences in n*:*

$$\begin{aligned}
r_1(n) &\equiv \max_{\bar{k} \in \{\hat{k}, \tilde{k}\}} \frac{E \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}{\sqrt{E \int L^*(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}} \\
r_2(n) &\equiv E \max_{k \in \{1, \ldots, K(n)\}} \sqrt{\int (L^{*0}_{n,S_n} - L^*)^2 (O, \psi_k(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)} \\
\tilde{r}(n)^2 &\equiv E d_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0).
\end{aligned}$$

Finally, for any $\delta > 0$ *we define*

$$\begin{aligned}
\epsilon_n(\delta) &\equiv (1+2\delta)\tilde{r}^2(n) + 2c(M_1, M_2, \delta) \frac{1 + \log(K(n))}{np} \\
&\quad + (1+\delta) r_1(n) \tilde{r}(n) + \frac{2M_3(1+\delta)\log(K(n))}{(np)^{0.5}} \max(r_2(n), (np)^{-0.5} I(r_2(n) > 0)).
\end{aligned}$$

Finite Sample Result. *For any $\delta > 0$, we have*

$$\begin{aligned} Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0) &\leq \left\{ \frac{r_1(n)(1+\delta) + \sqrt{r_1(n)^2(1+\delta)^2 + 4\epsilon_n(\delta)}}{2} \right\}^2 \\ &= f_1(\tilde{r}(n)^2, r_1(n), r_2(n), \log(K(n)), np, M_1, M_2, M_3(\log(K(n)), \delta). \end{aligned}$$

Define

$$\begin{aligned} &f_1(n,k) \equiv \\ &f_1(B_0(\epsilon_n(k)), r_1^*(n(1-p)), r_2^*(n(1-p)), \log(N(\epsilon_n(k))), n(1-p)p^*, M_1, M_2, M_3(N(\epsilon_n(k)), \delta). \end{aligned}$$

We have that

$$\tilde{r}(n)^2 \leq \tilde{r}(n)^{u2} \equiv \min_{k \in \{1,\ldots,K(n)\}} f_1(n,k).$$

If $\eta_n = \eta_0$ is known, then this finite sample bound reduces to (set $p^* = 1$):

$$\begin{aligned} Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0) &\leq \\ &(1+2\delta) \min_k \left\{ (1+2\delta) B_0(\epsilon_n(k)) + 2C(M_1, M_2, \delta) \frac{1 + \log(N(\epsilon_n(k)))}{n(1-p)} \right\} \\ &+2C(M_1, M_2, \delta) \frac{1 + \log(K(n))}{np}. \end{aligned}$$

Asymptotic Implication. *For any $\delta > 0$*

$$Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0) \leq (1+2\delta) Ed_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0) + O(H(n)),$$

where

$$\begin{aligned} H(n) \equiv \max \Bigg(& \frac{\log(K(n))}{np}, \frac{\log(K(n)) r_2(n)}{(np)^{0.5}}, r_1^2(n), r_1(n)\tilde{r}(n), r_1(n)^{1.5}\tilde{r}(n)^{0.5}, \\ & \frac{\sqrt{\log(K(n))} r_1(n)}{(np)^{0.5}}, \frac{\sqrt{\log(K(n)} r_2(n)^{0.5} r_1(n)}{(np)^{0.25}} \Bigg). \end{aligned}$$

Consequently, we have the following scenarios.
Optimal rate: *If $\max\left(\frac{\log(K(n))}{np}, r_1(n)^2, \log(K(n)) r_2(n)^2\right) = O(\tilde{r}(n)^2)$, then $H(n) = O(\tilde{r}(n)^2)$, and thus*

$$Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0) = O(\tilde{r}(n)^{u2})$$

Asymptotic equivalence: *If either* $\max\left(\frac{\log(K(n))}{np}, r_1(n)^2, \log(K(n))r_2(n)^2\right) = o(\tilde{r}(n)^2)$ *or* $\max\left(\frac{\log(K(n))^2}{np}, r_1(n)^2, r_2(n)^2\right) = o(\tilde{r}(n)^2)$, *then*

$$H(n) = o(\tilde{r}(n)^2).$$

Consequently, under these two possible scenarios under which $H(n) = o(\tilde{r}(n)^2)$, *we have*

$$\frac{Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0)}{Ed_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0)} \to 1 \text{ for } n \to \infty. \tag{40}$$

Finally, if these two possible scenarios hold with $\tilde{r}(n)$ *replaced by the random quantaty* $d_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0)$, *then*

$$\frac{d_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0)}{d_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0)} \to 1 \text{ in probability for } n \to \infty. \tag{41}$$

If $p = p_n \to 0$ *for* $n \to \infty$, $\max\left(\frac{\log(K(n))^2}{np}, r_1(n)^2, r_2(n)^2\right) = o(\tilde{r}(n)^2)$, *and*

$$\frac{E \min_{k \in \{1,\ldots,K(n)\}} \int L(O, \psi_k(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)}{E \min_{k \in \{1,\ldots,K(n)\}} E_{S_n} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)} \to 1,$$

then

$$\frac{E \int L(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)}{E \min_{k \in \{1,\ldots,K(n)\}} \int L(O, \psi_k(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)} \to 1 \text{ for } n \to \infty. \tag{42}$$

That is, the cross-validation selector $c_{\hat{k}}$ *for selecting the constant performs asymptotically exactly as well as the selector* $c_{\tilde{k}_n}$ *which for each given data set selects the optimal constant.*

Proof. We apply Theorem 1 to the candidate estimators $\psi_k(\cdot \mid P_n)$, $k = 1, \ldots, K(n)$. This shows that the above theorem is a direct application of Theorem 1. Regarding the finite sample bound, we note that

$$\begin{aligned}
\tilde{r}^2(n) &= E \int L(O, \psi_{\tilde{k}}(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\
&= E \min_{k \in \{1,\ldots,K(n)\}} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \\
&\leq \min_{k \in \{1,\ldots,K(n)\}} E \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O).
\end{aligned}$$

By the finite sample result for $\psi_{\epsilon_n(k)}(\cdot \mid P_n)$ established in the previous section, we can bound the expectation, conditional on S_n, as follows: for any $\delta > 0$

$$E \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O) \leq (1 + 2\delta) \times$$
$$f_1(B(\epsilon_n(k)), r_1^*(n(1-p)), r_2^*(n(1-p)), \log(N(\epsilon_n(k))), n(1-p)p^*, M_1, M_2, M_3(N(\epsilon_n(k)), \delta).$$

Since this bound does not depend on S_n, it also holds unconditionally. This proves the reported finite sample bound. □

15.2 General loss function.

Similarly, we obtain this theorem for loss functions not satisfying property A3.

Theorem 8 Assumptions.
A1. *The limit η_0 of η_n for $n \to \infty$ needs to be an element of $\Gamma(P_0)$.*
A2. *There exist a $M_1^* < \infty$ so that*

$$\sup_{\psi \in \mathbf{\Psi}} \sup_{O} L^*(O, \psi, \psi_0) \leq M_1^*,$$

where the supremum is taken over the support of the distribution P_0 of O.

Definitions. *We define the following constants: We define the following sequences in n:*

$$\begin{aligned} f(M_1^*, K(n), np) &\equiv 4M_1^{*2}\sqrt{\frac{\log K(n)}{np}} \\ r_1(n) &\equiv \max_{\bar{k} \in \{\hat{k}, \tilde{k}\}} \frac{E \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}{\sqrt{E \int L^*(O, \psi_{\bar{k}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}} \\ \tilde{r}(n)^2 &\equiv E d_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0). \end{aligned}$$

We also define

$$\epsilon_{1n} \equiv \tilde{r}^2(n) + f(M_1^*, K(n), np) + r_1(n)\tilde{r}(n).$$

Finite Sample Result. *We have*

$$\begin{aligned} E d_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0) &\leq \left\{ \frac{r_1(n) + \sqrt{r_1(n)^2 + 4\epsilon_{1n}}}{2} \right\}^2 \\ &= f_2(\tilde{r}(n)^2, r_1(n), \log(K(n)), np, M_1^*). \end{aligned}$$

Here $\tilde{r}(n)^2$ is bounded by:

$$\begin{aligned}\tilde{r}^2(n) &\leq \tilde{r}(n)^{u2} \\ &\equiv \min_{k\in\{1,\ldots,K(n)\}} f_2(B_0(\epsilon_n(k)), r_1^*(n(1-p)), \log(N(\epsilon_n(k))), n(1-p)p^*, M_1^*).\end{aligned}$$

If η_0 is known, then this finite sample bound reduces to:

$$\begin{aligned}Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0) &\leq \min_k \{B_0(\epsilon_n(k)) + f(M_1^*, N(\epsilon_n(k)), n(1-p))\} \\ &\quad + f(M_1^*, K(n), np).\end{aligned}$$

Asymptotic Implication. *We have*

$$Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0) \leq Ed_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0) + O(H(n)),$$

where

$$H(n) \equiv \max\left(\frac{\log^{0.5}(K(n))}{(np)^{0.5}}, r_1(n)\tilde{r}(n), r_1(n)^2, r_1(n)\frac{\log^{0.25}(K(n))}{(np)^{0.25}}, r_1(n)^{1.5}\tilde{r}(n)^{0.5}\right).$$

Optimal rate: *If* $\max\left(r_1(n)^2, \frac{\log^{0.5}(K(n))}{(np)^{0.5}}\right) = O(\tilde{r}(n)^2)$, *then* $H(n) = O(\tilde{r}(n)^2)$, *and thus*

$$Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0) = O\left(\tilde{r}(n)^{u2}\right).$$

Asymptotic Equivalence. *If* $\max(r_1(n), \log^{0.5}(K(n))/(np)^{0.5}) = o(\tilde{r}(n))$, *then then*

$$\frac{Ed_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0)}{Ed_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0)} \to 1 \text{ for } n \to \infty. \tag{43}$$

Finally, if $\max(r_1(n), \log^{0.5}(K(n))/(np)^{0.5} = o_P(d_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0))$, *then*

$$\frac{d_{n(1-p)}(\psi_{\hat{k}}(\cdot \mid P_n), \psi_0)}{d_{n(1-p)}(\psi_{\tilde{k}}(\cdot \mid P_n), \psi_0)} \to 1 \text{ in probability for } n \to \infty.$$

If $p = p_n \to 0$ for $n \to \infty$, (43) holds, and

$$\frac{E \min_{k\in\{1,\ldots,K(n)\}} \int L(O, \psi_k(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)}{E \min_{k\in\{1,\ldots,K(n)\}} E_{S_n} \int L(O, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)} \to 1,$$

then

$$\frac{E \int L(O, \psi_{\hat{k}}(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)}{E \min_{k\in\{1,\ldots,K(n)\}} \int L(O, \psi_k(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)} \to 1 \text{ for } n \to \infty.$$

15.3 Asymptotic implications for cross-validated epsilon-net estimator.

Thus, we have now shown that the estimator $\psi_{\hat{k}}(\cdot \mid P_n)$ is asymptotically equivalent with the estimator which for each given data set chooses the estimator among $\{\psi_{\epsilon_n(k)}(\cdot \mid P_n), k = 1, \ldots, K(n)\}$, which is closest to the true parameter ψ_0, where we know that each of the candidate estimators $\psi_{\epsilon_n(k)}(\cdot \mid P_n)$ itself converges to ψ_0 at a rate smaller than or equal to the non-adaptive optimal rate $r^*_{opt}(n)$. Consequently, our estimator does not only achieve at minimal the optimal rate of convergence, but it also selects the smoothing parameter ϵ so that it is optimal for the actual parameter value ψ_0. That is, the ϵ-selector adapts to underlying properties of ψ_0.

16 Finite sample result for adaptive cross-validated epsilon-net estimator.

We define the benchmark selector

$$(\tilde{\epsilon}, \tilde{s}) = \text{argmin}_{(s,\epsilon)\in\mathcal{A}_n} E_{S_n} \int L(O, \psi_{s,\epsilon}(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}) dP_0(O).$$

We also define

$$B_0(\epsilon, s) = \min_j \int L(O, \psi_j^{s,\epsilon} \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O).$$

16.1 Quadratic loss function

An application of Theorem 1 yields the following finite sample and asymptotic results for this estimator.

Theorem 9 Assumptions.
A1. *The limit η_0 of the nuisance parameter estimator $\eta_n = \eta(P_n)$ for $n \to \infty$ is an element of $\Gamma(P_0)$.*
A2. *There exist a $M_1^* < \infty$ so that*

$$\sup_{\psi\in\mathbf{\Psi}} \sup_O L^*(O, \psi, \psi_0) \leq M_1^*,$$

where the supremum is taken over a support of the distribution P_0 of O.
A3. *There exist a $M_2 < \infty$ so that for all $\psi \in \mathbf{\Psi}$*

$$\sup_{\psi \in \mathbf{\Psi}} \frac{VAR_{P_0} L^*(O, \psi, \psi_0)}{E_{P_0} L^*(O, \psi, \psi_0)} \leq M_2.$$

Definitions. *We define the following constants:*

$$\begin{aligned}
M_1 &= 2M_1^* \\
c(M_1, M_2, \delta) &= 2(1+\delta)^2 \left(\frac{M_1}{3} + \frac{M_2}{\delta} \right) \\
a_0 &\equiv 2M_1/3 \\
M_3(\log(K(n))) &= 3a_0 + \sqrt{2}\frac{\sqrt{\log(2)}}{\log(K(n))} + \frac{\sqrt{2}}{\sqrt{\log(K(n))}} + b_0 + \int_{b_0}^{\infty} 2K(n)^{1-m(x)} dx,
\end{aligned}$$

where b_0 is the smallest constant larger than the solution of $1 - m(x) = 0$ with $m(x) = 0.5\frac{x^2}{1/\log(K(n)) + a_0 x}$. We note that $M_3(\log(K(n)) \downarrow$ in n.

We also define the following sequences in n:

$$\begin{aligned}
r_1(n) &\equiv \max_{(\bar{s},\bar{\epsilon}) \in \{(\hat{s},\hat{\epsilon}),(\tilde{s},\tilde{\epsilon})\}} \frac{E \int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\bar{s},\bar{\epsilon}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}{\sqrt{E \int L^*(O, \psi_{\bar{s},\bar{\epsilon}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}} \\
r_2(n) &\equiv E \max_{(s,\epsilon) \in \mathcal{A}_n} \sqrt{\int (L^{*0}_{n,S_n} - L^*)^2 (O, \psi_{s,\epsilon}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)} \\
\tilde{r}(n)^2 &\equiv E d_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0).
\end{aligned}$$

Finally, for any $\delta > 0$ we define

$$\begin{aligned}
\epsilon_n(\delta) &\equiv (1+2\delta)\tilde{r}^2(n) + 2c(M_1, M_2, \delta)\frac{1 + \log(K(n))}{np} + \\
&\quad (1+\delta) r_1(n)\tilde{r}(n) + \frac{2M_3(1+\delta)\log(K(n))}{(np)^{0.5}} \max(r_2(n), (np)^{-0.5} I(r_2(n) > 0)).
\end{aligned}$$

Finite Sample Result. *For any $\delta > 0$, we have*

$$\begin{aligned}
E d_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0) &\leq \left\{ \frac{r_1(n)(1+\delta) + \sqrt{r_1(n)^2(1+\delta)^2 + 4\epsilon_n(\delta)}}{2} \right\}^2 \\
&\equiv f_1(\tilde{r}(n)^2, r_1(n), r_2(n), \log(K(n)), np, M_1, M_2, M_3(\log(K(n)), \delta).
\end{aligned}$$

Let

$$f_{1n}(s,\epsilon) \equiv$$
$$f_1(B_0(s,\epsilon), r_1^*(n(1-p)), r_2^*(n(1-p)), \log(N_s(\epsilon)), n(1-p)p^*, M_1, M_2, M_3(N_s(\epsilon)), \delta).$$

We have

$$\begin{aligned} \tilde{r}(n)^2 &\leq \tilde{r}(n)^{u2} \\ &\equiv \min_{(s,\epsilon)\in\mathcal{A}_n} f_{1n}(s,\epsilon). \end{aligned}$$

If η_0 is known, then this finite sample bound reduces to (set $p^* = 1$):

$$\begin{aligned} &Ed_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0) \leq \\ &(1+2\delta) \min_{(s,\epsilon)\in\mathcal{A}_n} \left\{ (1+2\delta)B_0(s,\epsilon) + 2C(M_1,M_2,\delta)\frac{1+\log(N_s(\epsilon))}{n(1-p)} \right\} \\ &\qquad\qquad +2C(M_1,M_2,\delta)\frac{1+\log(K(n))}{np}. \end{aligned}$$

Asymptotic Implication. *For any $\delta > 0$*

$$Ed_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0) \leq (1+2\delta)Ed_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0) + O(H(n)),$$

where

$$\begin{aligned} H(n) \equiv \max\Bigg(&\frac{\log(K(n))}{np}, \frac{\log(K(n))r_2(n)}{(np)^{0.5}}, r_1^2(n), r_1(n)\tilde{r}(n), r_1(n)^{1.5}\tilde{r}(n)^{0.5}, \\ &\frac{\sqrt{\log(K(n))}r_1(n)}{(np)^{0.5}}, \frac{\sqrt{\log(K(n)}r_2(n)^{0.5}r_1(n)}{(np)^{0.25}} \Bigg). \end{aligned}$$

Consequently, we have the following scenarios.
Optimal rate: *If* $\max\left(\frac{\log(K(n))}{np}, r_1(n)^2, \log(K(n))r_2(n)^2\right) = O(\tilde{r}(n)^2)$*, then* $H(n) = O(\tilde{r}(n)^2)$*, and thus*

$$Ed_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0) = O(\tilde{r}(n)^{u2}).$$

Asymptotic equivalence: *If either* $\max\left(\frac{\log(K(n))}{np}, r_1(n)^2, \log(K(n))r_2(n)^2\right) = o(\tilde{r}(n)^2)$ *or* $\max\left(\frac{\log(K(n))^2}{np}, r_1(n)^2, r_2(n)^2\right) = o(\tilde{r}(n)^2)$*, then*

$$H(n) = o(\tilde{r}(n)^2).$$

Consequently, under these two possible scenarios under which $H(n) = o(\tilde{r}(n)^2)$, *we have*

$$\frac{Ed_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0)}{Ed_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0)} \to 1 \text{ for } n \to \infty.$$

If these two possible scenarios hold with $\tilde{r}(n)$ *replaced by the random quantaty* $d_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0)$, *then*

$$\frac{d_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0)}{d_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0)} \to 1 \text{ in probability for } n \to \infty.$$

If $p = p_n \to 0$ *for* $n \to \infty$, $\max\left(\frac{\log(K(n))^2}{np}, r_1(n)^2, r_2(n)^2\right) = o(\tilde{r}(n)^2)$, *and*

$$\frac{E \min_{(s,\epsilon)\in\mathcal{A}_n} \int L(O, \psi_{s,\epsilon}(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)}{E \min_{(s,\epsilon)\in\mathcal{A}_n} E_{S_n} \int L(O, \psi_{s,\epsilon}(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)} \to 1,$$

then

$$\frac{E \int L(O, \psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)}{E \min_{(s,\epsilon)\in\mathcal{A}_n} \int L(O, \psi_{s,\epsilon}(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)} \to 1 \text{ for } n \to \infty.$$

That is, the final statements says that the cross-validation selector $(\hat{s}, \hat{\epsilon})$ for selecting the (s, ϵ) performs asymptotically exactly as well as the selector $(\tilde{s}, \tilde{\epsilon})$ which for each given data set selects the optimal (s, ϵ) making the estimator $\psi_{s,\epsilon}(\cdot \mid P_n)$ closest to ψ_0.

16.2 General loss function

Similarly, we obtain this theorem for loss functions not satisfying property A3.

Theorem 10 Assumptions.
A1. *The limit* η_0 *of* $\eta_n = \eta(P_n)$ *for* $n \to \infty$ *is an element of* $\Gamma(P_0)$.
A2. *There exist a* $M_1^* < \infty$ *so that*

$$\sup_{\psi\in\mathbf{\Psi}} \sup_{O} L^*(O, \psi, \psi_0) \leq M_1^* \text{ a.s.},$$

where the supremum is taken over the support of the distribution P_0 *of* O.

Definitions. *Let $B \equiv \{(\hat{s},\hat{\epsilon}),(\tilde{s},\tilde{\epsilon})\}$ be the collection of these two estimators. We define the following sequences in n:*

$$\begin{aligned} f(M_1^*, K(n), np) &\equiv 4M_1^{*2}\sqrt{\frac{\log K(n)}{np}} \\ r_1(n) &\equiv \max_{(\bar{s},\bar{\epsilon})\in B} \frac{E\int (L^{*0}_{n,S_n} - L^*)(O, \psi_{\bar{s},\bar{\epsilon}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}{\sqrt{E\int L^*(O, \psi_{\bar{s},\bar{\epsilon}}(\cdot \mid P^0_{n,S_n}), \psi_0) dP_0(O)}} \\ \tilde{r}(n)^2 &\equiv E d_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0). \end{aligned}$$

We also define

$$\epsilon_{1n} \equiv \tilde{r}^2(n) + f(M_1^*, K(n), np) + r_1(n)\tilde{r}(n).$$

Finite Sample Result. *We have*

$$\begin{aligned} E d_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0) &\leq \left\{ \frac{r_1(n) + \sqrt{r_1(n)^2 + 4\epsilon_{1n}}}{2} \right\}^2 \\ &= f_2(\tilde{r}(n)^2, r_1(n), \log(K(n)), np, M_1^*). \end{aligned}$$

Here $\tilde{r}(n)^2$ can be bounded as follows:

$$\begin{aligned} \tilde{r}^2(n) &\leq \tilde{r}(n)^{u2} \\ &\equiv \min_{(s,\epsilon)\in\mathcal{A}_n} f_2(B_0(\epsilon, s), r_1^*(n(1-p)), \log(N_s(\epsilon)), n(1-p)p^*, M_1^*). \end{aligned}$$

If η_0 is known, then this finite sample bound reduces to:

$$\begin{aligned} E d_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0) &\leq \min_{(s,\epsilon)\in\mathcal{A}_n} \{B_0(\epsilon, s) + f(M_1^*, N_s(\epsilon), n(1-p))\} \\ &\quad + f(M_1^*, K(n), np). \end{aligned}$$

Asymptotic Implication. *We have*

$$E d_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0) \leq E d_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0) + O(H(n)),$$

where

$$H(n) \equiv \max\left(\frac{\log^{0.5}(K(n))}{(np)^{0.5}}, r_1(n)\tilde{r}(n), r_1(n)^2, r_1(n)\frac{\log^{0.25}(K(n))}{(np)^{0.25}}, r_1(n)^{1.5}\tilde{r}(n)^{0.5} \right).$$

Optimal rate: *If* $\max\left(r_1(n)^2, \frac{\log^{0.5}(K(n))}{(np)^{0.5}}\right) = O(\tilde{r}(n)^2)$, *then* $H(n) = O(\tilde{r}(n)^2)$ *and thus*

$$Ed_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0) = O(\tilde{r}(n)^{u2}).$$

Asymptotic Equivalence. *If* $\max(r_1(n), \log^{0.5}(K(n))/(np)^{0.5}) = o(\tilde{r}(n))$, *then then*

$$\frac{Ed_{n(1-p)}(\psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P_n), \psi_0)}{Ed_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0)} \to 1 \text{ for } n \to \infty. \tag{44}$$

If $\max(r_1(n), \log^{0.5}(K(n))/(np)^{0.5}) = o_P(d_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0))$, *then*

$$\frac{d_{n(1-p)}(\psi_{\hat{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0)}{d_{n(1-p)}(\psi_{\tilde{s},\tilde{\epsilon}}(\cdot \mid P_n), \psi_0)} \to 1 \text{ in probability for } n \to \infty.$$

If $p = p_n \to 0$ *for* $n \to \infty$, *(44) holds, and*

$$\frac{E \min_{(s,\epsilon)\in\mathcal{A}_n} \int L(O, \psi_{s,\epsilon}(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)}{E \min_{(s,\epsilon)\in\mathcal{A}_n} E_{S_n} \int L(O, \psi_{s,\epsilon}(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)} \to 1,$$

then

$$\frac{E \int L(O, \psi_{\hat{s},\hat{\epsilon}}(\cdot \mid P^0_{n,S_n}) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)}{E \min_{(s,\epsilon)\in\mathcal{A}_n} \int L(O, \psi_{s,\epsilon}(\cdot \mid P_n) \mid \eta_0) - L(O, \psi_0 \mid \eta_0) dP_0(O)} \to 1 \text{ for } n \to \infty.$$

16.3 Asymptotic adaptivity.

Thus, if we control the number of tried values (ϵ, s) and the nuisance parameter η_0 can be estimated at a fast enough rate, then $d_{n(1-p)}(\hat{\psi}(\cdot \mid P_n), \psi_0)$ converges to zero at the same rate as $\min_{s,\epsilon} d_{n(1-p)}(\psi_{s,\epsilon}(\cdot \mid P_n), \psi_0)$. Suppose that, for each s, we choose values $\epsilon_n(k, s) = c_k \epsilon^*_{s,opt}(n)$, $k = 1, \ldots, K_1(n, s)$, which achieve the non-adaptive optimal rate of convergence for $\psi_{s,\epsilon_n(k,s)}$ for the subspace $\mathbf{\Psi}_s$. That is, if ψ_0 happens to be an element of $\mathbf{\Psi}_s$, then the estimator $\psi_{s,\epsilon_n(k,s)}$ would converge to ψ_0 at the optimal rate for the parameter space $\mathbf{\Psi}_s$. It follows that with this choice of ϵ-values, the cross-validated adaptive epsilon-net estimator $\hat{\psi}$ is minimax adaptive in the sense that it converges (w.r.t. $d_{n(1-p)}(\cdot, \cdot)$) to ψ_0 at the rate which is optimal for the smallest of the parameter spaces which still contains ψ_0.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Budapest: Academiai Kiado, 1973.

C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci.*, 99(10): 6562–6566, 2002.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

H. Bozdogan. Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse fisher information matrix. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*. Heidelberg: Springer Verlag, 1993.

H. Bozdogan. Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44:62–91, 2000.

L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754, 1992.

L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996a.

L. Breiman. Out-of-bag estimation. Technical report, Department of Statistics, U.C. Berkeley, 1996b.

L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–824, 1998.

L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984a.

L. Breiman, J.H. Friedman, R.A. Olshan, and C.J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks/Cole, Monterey, 1984b.

L. Breiman and P. Spector. Submodel selection and evaluation in regression. the x random case. *International Statistical Review*, 60:291–319, 1992.

P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76:503–514, 1989.

L.M. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986.

L.M. Le Cam and G. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag, New York, 1990.

R. B. Davis and J. R. Anderson. Exponential survival trees. *Statistics in Medicine*, 8:947–961, 1989.

D.L. Donoho. Le cam lecture: Estimation by ϵ-nets. *IMS Le Cam Lecture 2003 at the Joint Statistical Meeting, San Francisco*, 2003.

D.L. Donoho and I.M. Johnstone. Ideal denoising in an orthonormal basis chosen from a libary of basis. *C.R. Acad. Sci. Paris, Ser I*, 319:1317–1322, 1994.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328, 1975.

R.D. Gill, M.J. van der Laan, and J.M. Robins. Coarsening at random: characterizations, conjectures and counter-examples. In D.Y. Lin and T.R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–94, New York, 1997. Springer Verlag.

L. Gordon and R.A. Olshen. Tree-structured survival analysis. *Cancer Treatment Reports*, 69:10065–1069, 1985.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nanparametric Regression*. Springer-Verlag, New York, 2002.

L. Györfi, M. Kohler, A. Krzyźak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York, 2002.

M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96: 746–774, 2001.

W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1993.

W. Härdle and J. S. Marron. Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika*, 72:481–484, 1985a.

W. Härdle and J. S. Marron. Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, 13:1465–1481, 1985b.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

T. J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990a.

T.J. Hastie and R.J. Tibshirani. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46(4):1005–1016, 1990b.

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *Annals of statistics*, 19(4):2244–2253, December 1991.

M. Jacobsen and N. Keiding. Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics*, 23:774–86, 1995.

I.M. Johnstone. Oracle inequalities and nonparametric function estimation. *Journal der Deutchen Mathematiker Vereinigung, Proc. of the International Congress of Mathematicians, Berlin 1998*, III:267–278, 1998.

S. Keleş, M. van der Laan, and S. Dudoit. Asymptotically optimal model selection method for regression on censored outcomes. *Technical Report, Division of Biostatistics, UC Berkeley*, 2002.

C. Kooperberg, C.J. Stone, and Y. K. Truong. Hazard regression. *Journal of the American Statistical Association*, 90(429):78–94, 1995.

M. Leblanc and J. Crowley. Relative risk trees for censored data. *Biometrics*, 48:411–425, 1992.

M. Ledoux. On talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1996.

P. Massart. About the constants in talagrand's concentration inequalities for empirical processes. Technical report, Department of Mathematics, Paris-Sud, 1998.

M. Pavlic and M. J. van der Laan. Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics and Data Analysis*, 41:413–428, 2003.

B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, New York, 1996.

J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

J. Robins and A. Rotnitzky. *Recovery of information and adjustment for dependent censoring using surrogate markers*, chapter AIDS Epidemiology, Methodological issues. Bikhauser, 1992.

J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

J. M. Robins, A. Rotnitzky, and M.J. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W. van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

J.M. Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceeding of the Biopharmaceutical section*, pages 24–33. American Statistical Association, 1993.

E. F. Schuster and C. G. Gregory. On the non-consistency of maximum likelihood nonparametric density estimators. In W.F. Eddy, editor, *Computer Science and Statistics: Proceedings of the 13th Symposium on the interface*, pages 295–298, 1981.

G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464, 1978.

D. W. Scott and L. E. Factor. Monte carlo study of three data-based nonparametric density estimators. *Journal of the American Statistical Association*, 76:9–15, 1981.

M. R. Segal. Regression trees for censored data. *Biometrics*, 44:35–47, 1988.

J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993.

B. W. Silerman. A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *Journal of the American Statistical Association*, 79(387):584–589, 1984.

B. W. Silverman. *Density Estimation for Statistics and Data analysis*. Chapman & Hall, 1986.

P. Smyth. Model selection of probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10:63–72, 2000.

C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12:1285–1297, 1984.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147, 1974a.

M. Stone. Cross-validatory choice and assessment of statistics predictions. *Journal of the Royal Statistical Society, Series B*, 36(2):111–147, 1974b.

M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1): 29–35, 1977.

M. Talagrand. A new look at independence. *Ann. Probab.*, 24:1–34, 1996a.

M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563, 1996b.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. Technical report, Division of Biostatistics, UC Berkeley, 2003.

M.J. van der Laan and J.M. Robins. Unified methods for censored longitudinal data and causality. Springer, New York, 2002.

A.W. van der Vaart. A note on cross-validation theory articles by Sandrine Dudoit, Mark van der Laan, and Maarten Wegkamp. Technical report, Department of Statistics, Free University Amsterdam, 2003.

P. Zhang. Model selection via multifold cross-validation. *Annals of Statistics*, 21:299–313, 1993.